DFM Techniques for the Detection and Mitigation of Hotspots in Nanometer Technology

by

Kareem Madkour

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

With the continuous scaling down of dimensions in advanced technology nodes, process variations are getting worse for each new node. Process variations have a large influence on the quality and yield of the designed and manufactured circuits. There is a growing need for fast and efficient techniques to characterize and mitigate the effects of different sources of process variations on the design's performance and yield. In this thesis we have studied the various sources of systematic process variations and their effects on the circuit, and the various methodologies to combat systematic process variation in the design space. We developed abstract and accurate process variability models, that would model systematic intra-die variations. The models convert the variation in process into variation in electrical parameters of devices and hence variation in circuit performance (timing and leakage) without the need for circuit simulation. And as the analysis and mitigation techniques are studied in different levels of the design flow, we proposed a flow for combating the systematic process variation in nanometer CMOS technology. By calculating the effects of variability on the electrical performance of circuits we can gauge the importance of the accurate analysis and model-driven corrections. We presented an automated framework that allows the integration of circuit analysis with process variability modeling to optimize the computer intense process simulation steps and optimize the usage of variation mitigation techniques. And we used the results obtained from using this framework to develop a relation between layout regularity and resilience of the devices to process variation. We used these findings to develop a novel technique for fast detection of critical failures (hotspots) resulting from process variation. We showed that our approach is superior to other published techniques in both accuracy and predictability. Finally, we presented an automated method for fixing the lithography hotspots. Our method showed success rate of 99% in fixing hotspots.

Acknowledgements

I would like to take this opportunity to express my special gratitude to Professor Mohab Anis, for his endless support and excellent guidance. Mohab's patience, encouragement, and support have made the completion of this dissertation and my PhD studies possible. I am forever thankful for all what he did to me. I am thankful to my co-supervisor Professor Karim Karim for his continues help.

My committee members, Professor Vincent Gaudet, Professor Eihab M. Abdel-Rahman, Professor Andrei Sazonov and Professor Azadeh Davoodi, have provided helpful feedback throughout the process, which I greatly appreciate.

I highly appreciate Sarah Mohamed for her constructive discussions and helpful comments. Her suggestions led to a significant improvement of this thesis.

I also would like to extend my appreciation to my current and past colleagues in Mentor Graphics, Andres Torres, Esraa Swillam, Fedor Pikus, Salma Mostafa, Marwa Shafee, Wael Manhawy, Dina Tantawy, Asmaa Rabee and Sherif Hammouda for their discussions and feedbacks. I greatly enjoyed working with intelligent, thoughtful and accomplished people.

I am grateful to Mentor Graphics for providing financial funding for my degree. Thanks to Hazem El-Tahawy and Juan Rey for making this possible.

I would also like to express my deepest gratitude to my family. My beloved wife, Abir ElMouelhy, for her love and endless support. My parents, Madkour Bakr and Aida AbdelHafiz for their unconditional love and support throughout my life. My brother, Diaa Madkour for his continuous encouragement.

Dedication

To my ever supportive and loving family.
Thank you.

Table of Contents

List of Tables

List of Figures

Nomenclature

ACLV Across-chip linewidth variation

ANN Artificial neural network

ASIC Application-specific integrated circuit

BEOL Back-end of line

CAD Computer aided design

CD Critical Dimension

CMOS Complementary metal oxide semiconductor

CMP Chemical-mechanical polishing

DFM Design for manufacturability

DIBL drain-induced barrier lowering

DP Double patterning

DRC Design rule checks

DSA Direct self assembly

DSL Dual stress liner

DSM Deep sub-micron

DUV Deep Ultra-violet

EPE Edge-placement error

EUV Extreme Ultra-violet

FEOL Front-end of line

FET Field effect transistor

FOCSI Fixed origin corner Square inspection

HCI Hot Carrier Injection

HiK+MG High-K metal gate

HP Half pitch

IC Integrated circuits

IP Intellectual property

ITRS International Technology Roadmap for Semiconductors

LER Line-edge roughness

LPE Layout physical extraction

MBE Model-based extraction

MEF Mask error factor

ML Machine learning

NA Numerical aperture

NBTI Negative-bias temperature instability

NRG Non-rectangular gate

OAI Off-axial illumination

OPC Optical Proximity Correction

PDK Process-design kit

PM Patterns matching

PnR Place-and-Route

RDF Random dopant fluctuation

RDR Restricted design rules

RET Resolution enhancement techniques

RIE Reactive ion etching

RMS Root Mean Square

RTA Rapid thermal annealing

SMT Stress memorization technique

SOI Silicon-On-Insulator

SRAF Sub-resolution assist features

STA Static timing analysis

STI Shallow trench isolation

SVM Support Vector Machine

VLSI Very large scale integration

WC-BC Worst case - best case

Chapter 1

Introduction and Motivation

1.1 Challenges

Process variations result from manufacturing imperfections. These variations increasingly affect the reliability, the functional yield, and the performance of modern CMOS processes. As CMOS feature size scales downward, the interaction between design and process intensifies and the gap between the expected and manufactured characteristics widens further, causing a significant impact on the chip yield and reliability.

Significant efforts in the semiconductor industry have started to deploy new tools and methodologies commonly referred to as DFM (Design for Manufacturability), to combat the effects of process variations. These DFM tools identify and correct the locations in a design where particle defects or process variations can create shorts and opens that cause functional failures. These faulty locations are called hotspots. DFM methodologies are essential to the development of new process, especially in the early development phases; when low functional yield is the primary obstacle to process qualification, yet they are still crucial to the most important milestone for the design: meeting the power and timing specifications. Traditional DFM techniques are essentially geometric operations with limited electrical interactions or awareness. These include resolution enhancement techniques to improve fidelity of optical lithography, design rule checks to restrict the use of layout

patterns not amenable to manufacturing, and guard-banding to keep margins for process variability in design. As the extent and complexity of process variations increase, and sub optimality due to conservative design threatens to offset the benefits of scaling, these traditional DFM techniques, while still crucial, are no longer sufficient.

Effective DFM techniques for mitigation of process variations require both understanding and characterization of the process variations effects at all levels of design. It is challenging for a designer to account for process variations correctly and decide which is the optimum decision for variations-tolerant design. This will require a huge amount of process development knowledge, device physics and circuit knowledge that are not common to find in a single organization not to mention in a single person. The growing trend for companies to adopt a fab-less business model limits the ability for design teams to interact with process engineers making it more difficult for designers to react toward variability. Moreover, different designs react differently to process variations according to their function, their design style and their constraints. Noting that not all the circuits in the design have the same effect in terms of performance and power. For example a slight variation in the arrival time of a signal in a path with large slack is not as important as the variation of the arrival time of a signal in a critical path. Any solution that treats the variability in these two paths equally will either be over-constraining or sensitive to variability.

1.2 Motivation

Many DFM techniques are applied to mitigate the impact of process variations. These techniques can be characterized as pure process techniques, process-design co-optimization techniques, and pure design techniques. In this work we will be focusing on mitigation techniques that can be done in the design side that can help the circuits meet their specifications, especially in the regime of stringent timing and power budgets, and detection and fixing catastrophic failures. The semiconductor industry needs to overcome the challenges listed in the previous section in order to be able to continue to march toward a successful business. Without the proper tools to guide the circuit designers, there is no guarantee that the fabricated circuits will match their specifications. Our attempt is to provide an

automated framework that can automatically characterize and mitigate the effects of different sources of process variations on different level of designs. The framework uses an acceptable level of abstraction; it should hide the details of the process from the designers and only guide the designers to mitigate the negative impact of variability only where the variability impact will affect the circuit performance or cause a systematic failure. This should positively reduce the design-to-market cycle and guarantee the acceptable level of yield.

1.3 Thesis Structure and Contribution

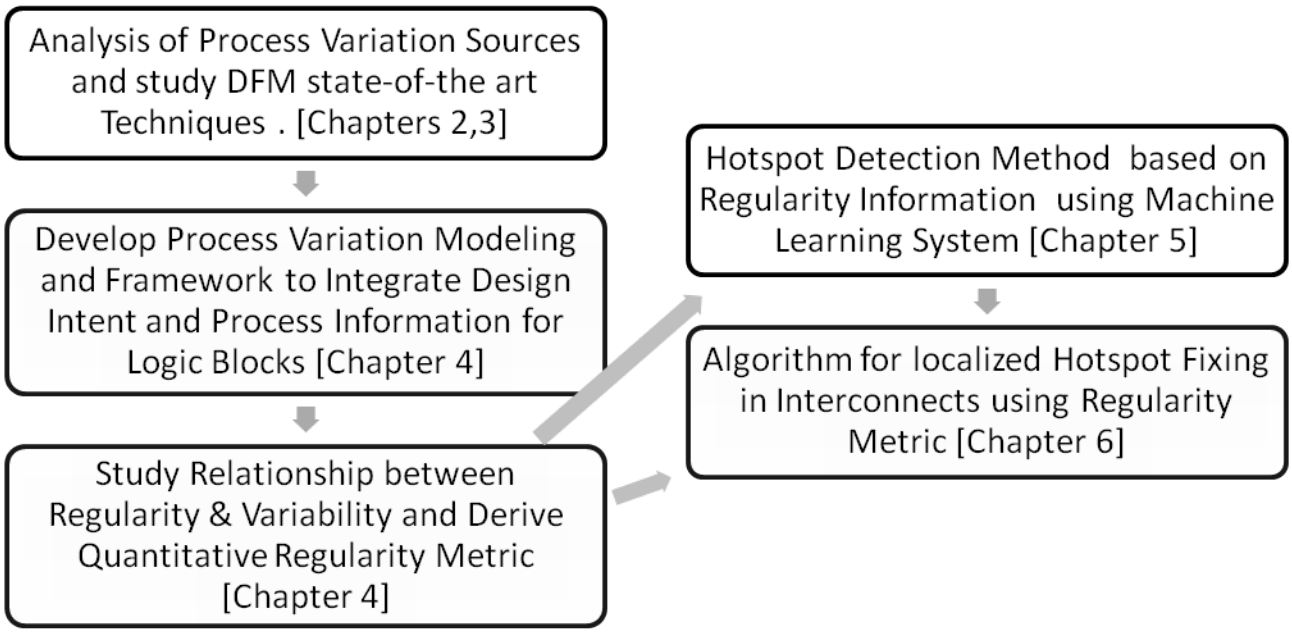


Figure 1.1: Thesis Structure and Contribution

This work will propose a flow to overcome the systematic process variations in nanometer CMOS technology. By studying the effects of variability on the electrical performance of circuits we can gauge the importance of the accurate analysis and model-driven corrections. Methods to improve robustness of circuits using these observations will be suggested. The proposed flow will be built on an automated framework and will be used in the design flow. The flow will use an acceptable level of abstraction that can automatically characterize and mitigate the effects of different sources of process variations on different level of designs. The proposed framework will hide the details of the process from the designers and guide the designers to mitigate the negative impact of variability only where circuit timing performance, power budget or yield are affected. Such a framework will provide the designers with a simple methodology to analyze the effects of variability, and also it will automatically fix the variability problems. This will positively reduce the design-to-market cycle and guarantee the acceptable level of yield.

Our contributions in this thesis are shown in figure 1.1 and can be listed as:

1. We will be presenting abstract and accurate process variability models that would model systematic within-die lithographic process variations. Unlike traditional methods that extract effective device dimensions and pass them to circuit simulators, our proposed models directly convert the variation in process into variation in electrical parameters of devices and hence variation in circuit performance (timing and leakage) without the need for circuit simulation.

2. An automated framework that allows the integration of circuit analysis with process variability modeling to optimize the computer intense process simulation steps and optimize the usage of variation mitigation techniques is implemented and its application is demonstrated.

3. We conduct an analysis of the causes of variation in the electrical parameters of devices and developing a relation between layout regularity and resilience of the devices to process variation. And we develop a fast quantitative metric to measure the layout regularity, this regularity metric will be used to in detecting problematic areas in the designs and select the optimum fixing.

4. We develop a novel technique for fast detection of critical failures (hotspots) resulting from process variation. The technique adopts support vector machine (SVM) as supervised machine learning approach. Data clustering and data balancing techniques are used to enhance the accuracy of the system. This technique is valid for detecting critical failures in front-end and back-end layers.

5. We develop an automated technique for fixing hotspots by performing localized changes in the design to improve the regularity and the robustness of the design against process variations, while following the design rules and preserving the circuit connectivity.

The items (1-3) in the contribution list will be covered in chapter 4, item 4 will be covered in chapter 5 and item 5 will be covered in chapter 6. The remaining structure of the thesis is as follows:

- In Chapter 2, we will review process variations sources and effects. We will classify variations according to their spatial scale, cause and behavior. Then we will review the different sources of variations along the different fabrication steps. We will show how systematic within-die variations are dominant in new technology nodes. We will also review the effects of variations on the transistor level and the circuit performance.

- Next, in Chapter 3, we will review different DFM techniques for characterizing systematic variations and mitigating them in the design phase. We will start by comparing statistical design methods and model based design DFM approach. We will review the published approaches for variation-aware analysis and design along the various stages of the design from synthesis, placement, routing, layout, post-tapeout fixing.

- We will present modeling of CMOS transistor electrical parameters variation resulting from lithography and stress effects in Chapter 4. Then we will propose a new DFM framework that provides an automated analysis and mitigating solution to the

problem of process variations. The relation between electrical variability and regularity of the design is studied. A fast quantitative metric to measure the regularity of designs is derived and compared against modeling of process variation effects.

- In Chapter 5, we will present a novel technique for detection of critical catastrophic failures in the design. The approach will use the relation between variation and regularity derived in 4 and will utilize an SVM (support vector machine) classifier to detect the failures. Several techniques, including topological data clustering and data balancing are provided to enhance performance of the proposed approach. We will show that our approach is superior to other published techniques in both accuracy and predictability.

- Chapter 6 will present an automated method for fixing the lithography hotspots. The method will integrate the regularity metric developed in 4. We will apply the fixes on real 32nm design and validate that the fixes provided are both design-rules violations clean and lithography failures clean.

- Chapter 7 is the final summary of the work performed for this thesis and the presentation of possible future directions.

Chapter 2

Background: Process Variations

2.1 Historical Overview of Process Variations

In 1965, Gordon Moore was the first to observe that the cost of integrated circuits (ICs) was minimized by doubling the number of components (transistors) on IC every year [1]. (Although originally calculated as a doubling every year, Moore later refined the period to two years, and currently it is often quoted as every 18 months [2].) Half a century later this financial remark is still ruling the semiconductor industry. Achieving Moore's law (prophecy) of cost scaling was made possible by a continuous innovation in the fields of materials, devices and mostly optical lithography through physical scaling down of the CMOS features.

While the continued decrease in the ratio of feature sizes to fundamental dimensions (such as atomic dimensions and light wavelengths) means that management of variations will play a significant role in future technology scaling, the evidence shows that process variations have been a continuing theme throughout semiconductor history. Though process variations are sometimes treated as new challenges associated with technology scaling, the problem of variations has been studied for almost 50 years. In 1961, Shockley analyzed the random fluctuation in junction breakdown [3]. Systematic variations in MOS devices were first addressed formally in 1974 by Schemmert and Zimmer [4] when they computed

the sensitivity of ion-implanted MOS threshold voltages as a function of the implantation energy and the oxide thickness.

Process variations have always been a critical problem in semiconductor fabrication and understanding and mitigating process variations has been a continuing target for semiconductor process engineers. Starting from around the 180nm CMOS technology node, variations started to be a big source of concern for the design community. At this node, the patterns lithographically "printed" on silicon were for the first time smaller than the wavelength of light (193nm) patterning them. This results in sub-wavelength optical phenomena that introduces undesirable effects. Clever process tricks, as well as good modeling of these effects, were able to neutralize large parameter uncertainties for another couple of technology nodes, at which point variations were no longer simply wafer-to-wafer or die-to-die shifts.

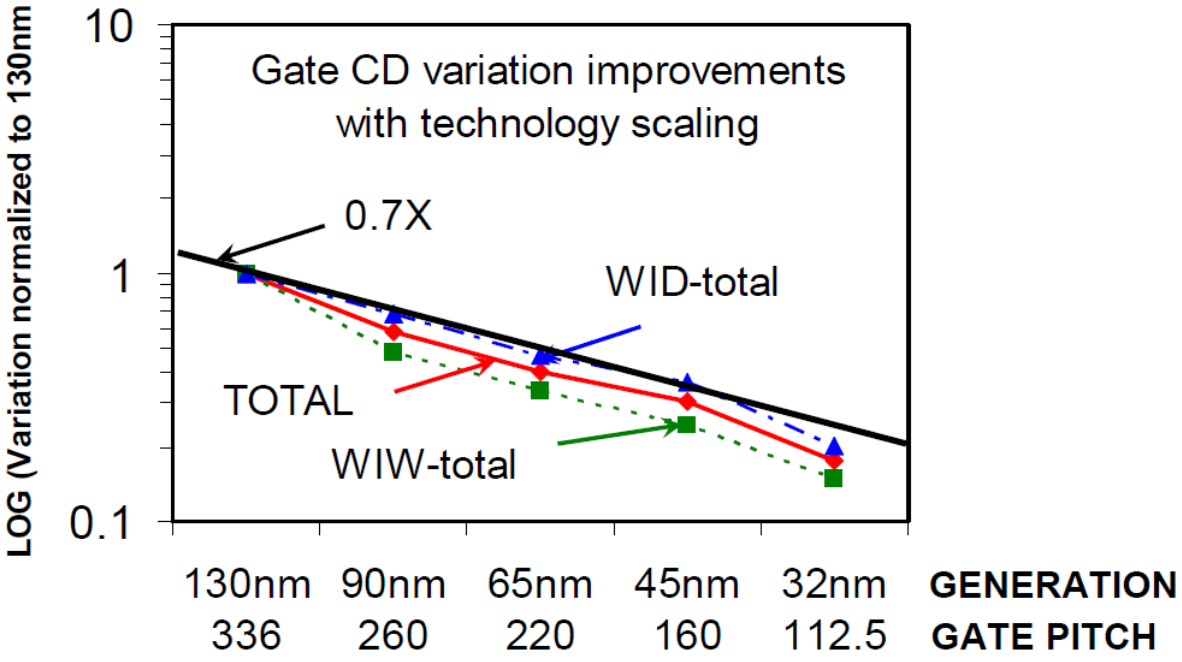


Figure 2.1: Gate CD Variations across technology nodes [5].

Instead, the variations problem began to significantly affect circuits within-die [6]. Fig-

ure 2.1 shows the trend of transistor gate critical dimension variation across various technology nodes in Intel [5]. It is shown how Intel succeeded in scaling down the variation with the same order as technology scale (0.7X) but the figure also shows how within-die variations are becoming more significant contributors to the total variation. At 45 nm, the systematic variation between two transistors of the same L and W can reach more than 30% in drive current (ION) and as high as 100 mV in threshold voltage (VTH) [7]. This resulted in a flurry of research on the underlying causes of variations, efficient modeling methods to enable designers to study the effects during design, changes or additions to the manufacturing process to reduce absolute uncertainties, and circuit techniques to moderate the impact to sensitive circuits. This situation will continue to deteriorate rapidly due to the increase in relying on the Deep Ultra Violet (DUV 193nm-Technology) that is operating at beyond its originally intended limits.

2.2 Classification of Process Variations

One can describe the variability mechanism from different perspectives. Variations can be classified from the perspective of :

- spatial scale.
- behavior (Random & Systematic).

2.2.1 Spatial Scale of Variations

The first categorization of variations will be on spatial scale, we will attempt to describe how variation sources manifest themselves on different spatial scales. The variations can be separated into lot-to-lot, wafer-to-wafer, inter-die, and intra-die variations. Lot-to-lot, wafer-to-wafer, and inter-die variations are because of the differences in manufacturing conditions between lots, wafers and dies. It is the challenge of process engineers to ensure better control and uniformity during the manufacturing process. Die-to-die variations occurs when the same transistor fabricated on different dies have different electrical properties.

These are variations between identically designed structures on different dies. Inter-die variations can cause shifts in performance between dies. Within-die (intra-die) variations occur when nominally identical transistors on the same die have different electrical properties after fabrication. Intra-die variations can introduce significant offset and matching problems between transistors. While matching has long been a problem in analog circuit design, recently, digital circuit designers have also begun to worry about intra-die variations. For example, within-die variations in gate critical dimension (CD) can affect the circuit performance and in the worst case, variations can even result in non-functioning designs, significantly decreasing the yield of the circuit.

2.2.2 Systematic and Random Behavior of Variations

There are two types of variations behavior: systematic behavior and random behavior. By definition, systematic variations describe the behavior of variations that can be analyzed in a methodical way, and can be formulated by function and its effect is calculated. Therefore, systematic variations are also called deterministic variations. Accurate modeling of systematic variations of the design can be used to predict design behavior. Designers can take advantage of this predictability to design circuits accordingly and avoid designing for the worst-case. For example, gate critical dimension vary according to the spacing to neighboring gates. A designer can either use a predictive model to estimate how much variations will be expected from the layout, or reduce the systematic variations by inserting the polysilicon at a regular spacing to maintain a uniform pitch and hence a well predicted dimension.

Another class of behavior falls into the category of random variations. Random variations are variations for which the designers do not have enough information to quantitatively or functionally relate to its origin of variations, and therefore are forced to design for the worse case. This means a large design margin must be incorporated to compensate for the worst case scenario. Designing for the worst case can waste resources that can potentially be used for performance improvement or energy reduction. Every effort should be made to understand this kind of variations better in order to minimize the design cost associated with accommodating it. Variations that are referred to as systematic relies solely

on the fact that designers can trace the origin of the variations back to a specific design parameter. In other words, nothing prevents a random variation source from becoming a systematic variation source if researchers can find a way to relate the variation source to a specific design parameter.

Ring oscillators are useful tools to measure random and systematic variations. Measuring the variation of oscillation frequency of closely spaced ring oscillators are used to obtain random variation data, and large populations of oscillators (with random variation removed via RMS) are used to obtain systematic variation data. The variability breakdown described in [8] is extremely important. They analyzed a large set of data from IBM's 65nm SOI technology, comprising 23 lots with 24 wafers each, approximately 100 die per wafer, and 14 ring oscillators across each die. The total number of dies (excluding some missing points) was about 36000. The data shows that 70% of the total variability is from one die to another, 20% is due to systematic within-die effects, and only 10% is due to random or unknown sources of variations. This shows that the bulk of design variability can be captured using traditional corner-based analysis. But the importance of the modeling of the systematic with-in die variations is that these are the variations that designers with the aid of accurate models can null.

Figure 2.2 shows how Intel managed to reduce the effects of random variations starting from 45nm technology with the introduction of high-K metal gate technology, while the systematic variations remained constant resulting in increase in the contribution of systematic variations to the total variations. [9] We have already showed in Figure 2.1 how the percentage of with-in die variations is increasing in the total variation budget. For these reasons we will focus our work on analyzing and fixing effects of systematic with-in die variations.

2.3 Sources of Variations

In Deep-Sub-Micron (DSM) CMOS, each of the fabrication step requires one or more unit process steps. For example, formation of one of the two twin-well implants involves depositing or thermally growing an oxide layer, spinning on photoresist, lithographically

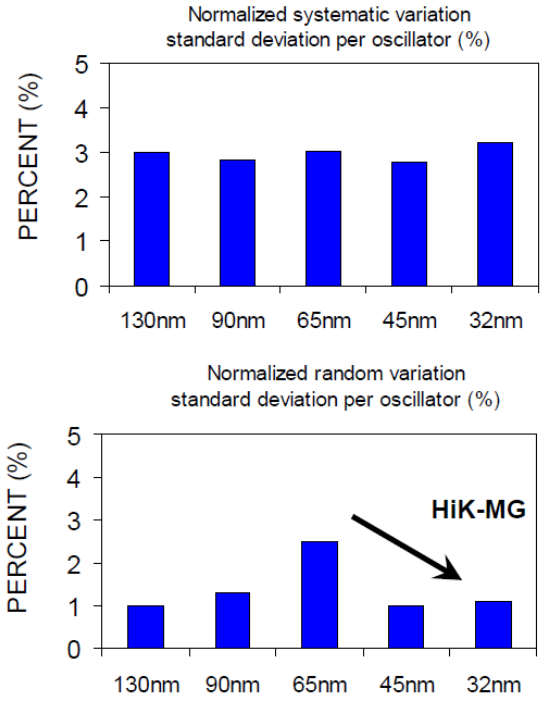


Figure 2.2: Systematic and Random Variation Technology trends. [9]

patterning the photoresist to define the well area, developing away the exposed photoresist over the defined well areas, implanting the appropriate dopant species and then removing the resist and oxide layer. This entire procedure is then repeated for the other well.

Variations occur in all of the steps of manufacturing the chip. However, a number of these processing steps can be highlighted as major sources of variations: (1) photolithography, (2) ion implantation and thermal annealing, (3) etch, (4) shallow trench isolation (STI) and sidewall spacer stress, and (5) chemical-mechanical polishing (CMP). Depending on the features being fabricated, each processing step affects the circuit differently. Photolithography, etch and CMP variations would affect the physical fabricated dimensions of transistors and tracks, while ion implantation, thermal annealing and lattice stress would influence the internal molecular composition of the material making up the transistors. Figure 2.3 summarizes the different variation sources. In the section below, we will iden-

tify the processing steps which most induce variations and also point out the transistor parameters that are most affected by them.

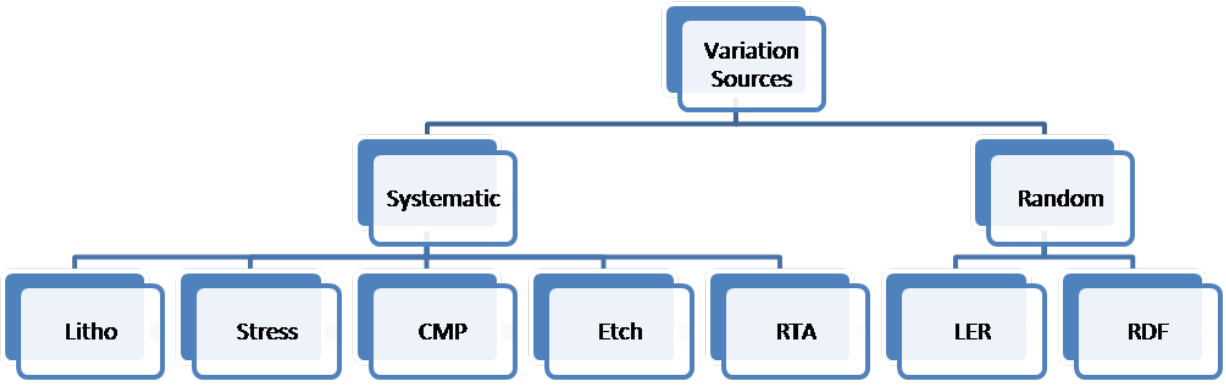


Figure 2.3: Sources of Variation

2.3.1 Photolithography

The photolithography process is used to project the design pattern from photo-masks onto the actual wafer. A simplified view of a modern lithography system is shown in Figure 2.4. The illumination source (laser) produces a coherent light wave of a certain wavelength, currently 193nm. The light illuminates the mask containing the layout pattern. The light diffracts from mask openings and forms a series of diffracted beams at a finite number of angles that are dependent on the wavelength. The diffracted beams that pass through the lens are combined at the wafer (image plane) and form an interference pattern. In an ideal situation, it is best to use a light wavelength that is equal or shorter than the critical dimension (CD) in that technology.

As we continue to scale sub 100nm, the lithography process cannot keep up with the aggressiveness of scaling, and newer technologies continue to base their lithography pro-

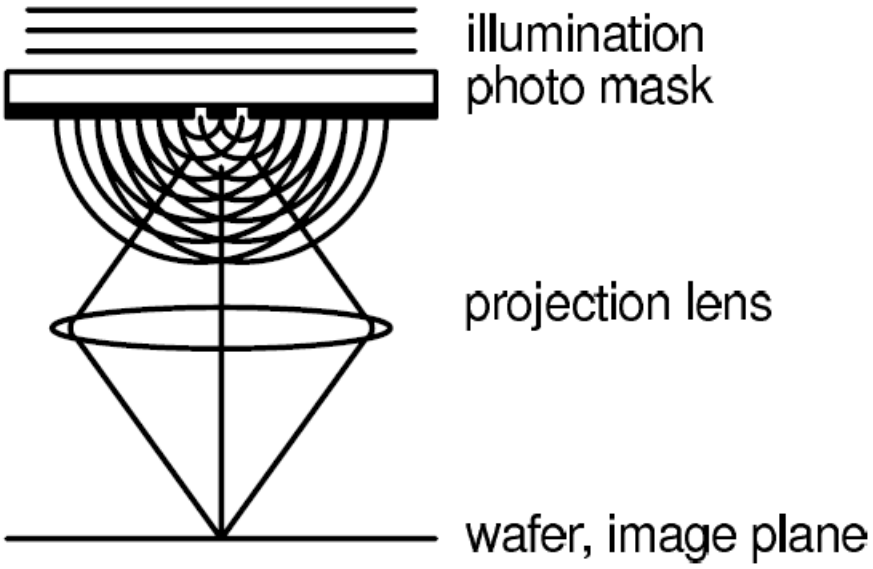


Figure 2.4: Simplified view of the illumination system.

cess on wavelengths that are much longer than the CD. The minimum printable pitch is governed by Rayleigh's resolution criteria that directly relates the resolution limit to the optical wavelength used in lithography.

$$HP = K_1 \frac{\lambda}{NA}, \tag{2.1}$$

where HP, λ, and NA represent the half pitch (critical dimension), the wavelength, and the numerical aperture of optical system, respectively. The parameter K_1 depends on the process specifications. The complexity of a lithography process is graded in terms of the empirically determined K_1 factor defined in the above equation. A smaller k1 factor indicates that the lithography process can resolve a smaller half-pitch for the same wavelength and numerical aperture (NA), which is often achieved through an increase in cost and complexity and leads to degrading pattern fidelity. Shape distortion can be attributed to the low-pass filter behavior of the lithography process while trying to print smaller features than the light wavelengths. The low-pass filter characteristics can result in inaccuracy while resolving the high frequency components, such as corners or sharp turns on the wafer. This inaccuracy translates into several major types of distortions: line-width

variations (proximity effect), line-end shortening and corner rounding. Due to the strong layout dependence, these kinds of variations are highly systematic. The proximity effect refers to the strong dependence of the printed critical dimension on the surrounding layout. The closer the surrounding layout is, the more impact it is going to have on the printed dimension of the transistors around it.

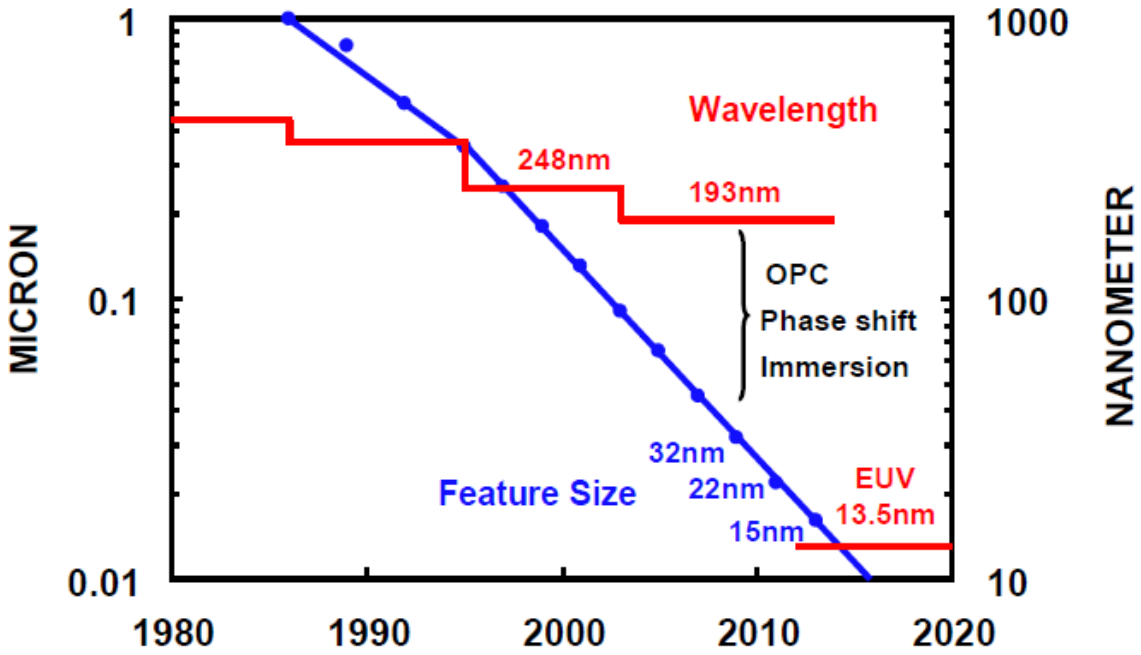


Figure 2.5: Minimum allowable pitch with wavelength for each technology node. [9]

As we scale from one generation to another, the minimum allowable feature size is getting smaller as shown in Figure 2.5. Meanwhile, the lithography wavelength have not decreased beyond 193nm, creating a large "'technology gap"' that increases with the march toward smaller technologies. And with the delay of the much anticipated for EUV technology, one has to expect more challenges to realize smaller technologies. As long as the optical lithography wavelength remains constant and the minimum allowable feature size continues to shrink, therefore, more interaction between neighboring transistors is expected as the technology nodes advance. That is one reason why scaling has increased the prob-

lem of variations. Line shortening refers to the reduction in line length while printing a rectangular structure. This is due to both the diffraction of light and photoresist diffusion. Corner rounding refers to the smoothing of a rectangular corner into a rounded corner, mainly due to the low-pass filter characteristics of the lithography process. Moreover unintended changes in the photolithography parameters increases the degradation of the imaging process. The most influential parameters of lithography are: exposure energy, (de)focus plane change, mask bias, mask misalignment and flare.

The advance in photo-resist technology leads to resist materials with very high sensitivity. The kinetics of the resist is dependent on the sensitivity of the resist material and the light energy falling on the resist. A slight change in the exposure energy (often referred to dose) would have an effect on the feature dimensions. Defocus is defined as the distance, which is measured along the optical axis (i.e., perpendicular to the plane of the best focus) between the position of a resist-coated wafer and the position if the wafer was at the best focus. Vertical displacements due to non-uniform topography of photoresist during exposure will lead to the change in the image plane. A variety of factors such as wafer topography, CMP-driven layer non-uniformity, non-flatness of the mask, and focus setting error lead to the lack of focus, or de-focus, in imaging of patterns. The nominal image intensity profile is defined as being in-focus and printed with the nominal intensity dose. The amount of defocus determines the deviation of the printed geometry from the nominal geometry. Many patterns exhibit very high sensitivity to defocus. The defocus results in blurring of the image transferred onto the wafer and consequently translates to the line-width variations.

The extent of variations depends on the line pitch. The Bossung plot in Figure 2.6 shows the variations of printed line-width at different pitch and defocus conditions. We observe that dense lines tend to smile with defocus, whereas isolated lines frown.

The error in mask making is becoming more important as we proceed to scale down the mask dimensions. For large features a unit change in the mask dimension corresponds to a unit change in the wafer dimension scaled by the demagnification of the exposure system. This is an important advantage of reduction projection systems. For example, a 40 nm mask dimensional error results in only a 10 nm line-width error for a 4X system. In low K1 imaging, however, the benefit due to this demagnification is reduced. In a loose sense,

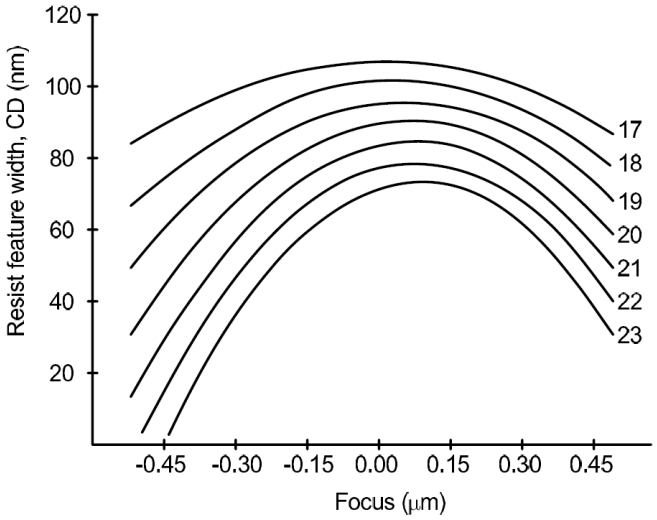


Figure 2.6: Dependence of line-width on defocus for patterns with different pitches [10].

mask dimensional error is magnified. The severity of such amplification is described by the mask error factor (MEF), With a MEF of 2.0 on a 4X system, for instance, a 40 nm mask dimensional error results in a 20 nm rather than 10 nm resist line-width error!

Controlling the alignment between different masks is important to prevent circuit failure (e.g., contact-metal misalignment) and also in critical dimension control. In single exposure systems, as shown in Figure 2.7 any misalignment between diffusion and polysilicon will result in change in the transistor's gate width and length. And with the introduction of double patterning technology, overlay error is now a major problem for critical dimension uniformity [11].

Another type of lithography variations is caused by random uncertainties in the fabrication process is Line-Edge Roughness (LER). LER is mainly caused by erosion of polymer aggregates at the edge of photo-resist (PR) during development and fully depends on some complex chemical formula, it is so difficult to generate the LER image in print-images of layouts, and in our knowledge no commercial lithography simulation tools can generate print images caused by LER. Even though LER is a kind of random variations, it is undesirable and has to be analyzed because it highly degrades the device performance. LER is

Figure 2.7: Layout view (left) and simulated post-lithography image (right) of a device.

on the order of several nanometers, and can be one of the performance limiting components for 45nm and below technologies.

2.3.2 Etching

Similar to the photolithography process, the etching process has non-uniformities which also contribute to the line-width variations. We can classify etch induced variations into three categories [12] as shown in Figure 2.8.

The first group is composed of etch rate and profile deviations that are caused by kinetic ion and neutral fluxes. Angular dispersion of ions and neutrals due to collisions within the sheath, and ion and neutral interaction with sidewalls result in positive RIE-lag, negative RIE-lag, faceting, micro-trenching, retrograde sidewall, and sloped sidewalls. The second group is composed of etch profile deviations from design that are induced by electron charging of the wafer substrate. The electron charging alters the trajectory of high energy ions while they are in transit through the micro-structure. This electron shading effect is caused by non-uniform charging of the etching feature; upper parts of the feature and its sidewalls are locally charged and can deflect the flight of the ions reaching the bottom of the micro-structure. Electron charging reduces the number of the etching species reaching

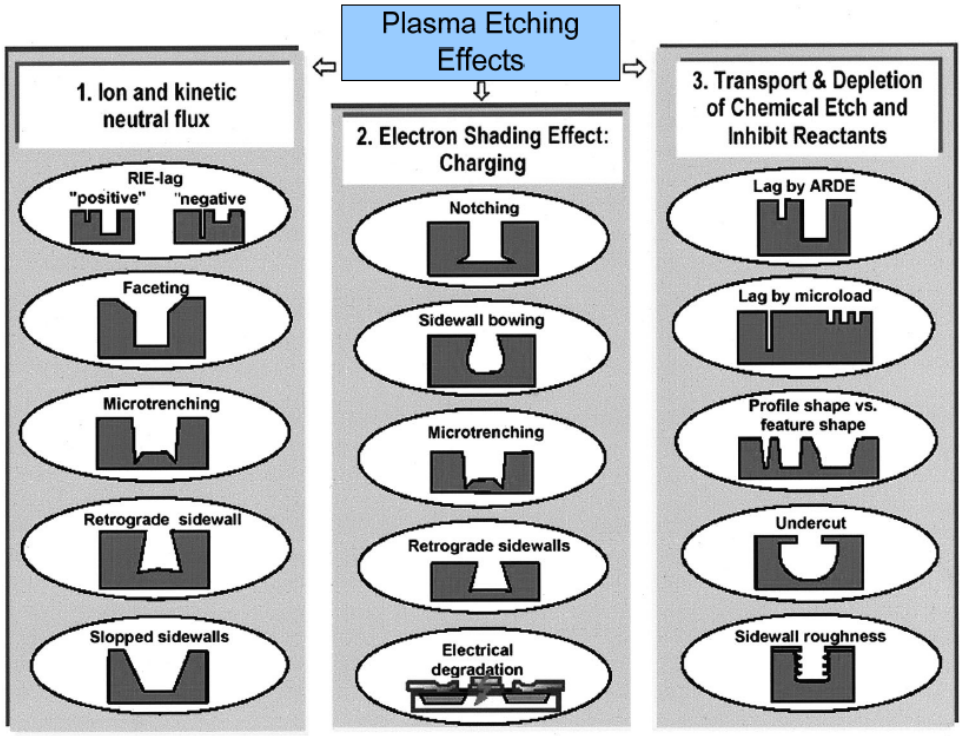


Figure 2.8: Plasma etch non-uniformity effects [12]

the bottom surface. Pattern dependent charging originates in the directionality differences between ions and electrons as they cross the plasma sheath and interact with both conducting and insulating micro-structures. The results of surface charging are notching, sidewall bowing, micro-trenching, and electrical degradation or plasma damage. The third group is composed of microscopic non-uniformities due to transport and depletion of chemical etch and inhibitor reactants. Ion and neutral transport and depletion cause both etch rate and profile deviations, and are significant and dominant factors in microscopic non-uniformity, RIE-lag, and micro-loading. Furthermore, transport and depletion of chemical etching and inhibitor reactants under conditions of high reaction probability at the wafer surface, along with deposition of material produced in discharge within the micro-structure, result in RIE-lag, micro-loading, irregular feature shape, undercutting, and sidewall roughness.

2.3.3 Ion Implantation and Thermal Annealing

During ion implantation, nuclear collisions during high energy implantation cause substrate atoms to be displaced, leading to materials defects. Rapid thermal annealing (RTA) is a subsequent annealing process, carried out at high temperature for a short duration, that is used to solve this problem. During RTA, the wafer is heated to around 1000^oC for a brief period and then cooled. The RTA process thermally vibrates the atoms, reforms the bonds among them, and activates the dopants. The basic mechanism of RTA is to use radiation to rapidly transfer large heat flux to the wafer surface; this heat then spreads around the silicon wafer by conduction. Therefore, the surface temperature is due to both radiation and conduction. However, in today's RTA process, the time duration of heating is so short (less than 1 second) that complete thermal equilibrium between conduction and radiation cannot be achieved, and the surface temperature is primarily determined by the ability of different regions of the wafer to absorb heat from the RTA lamps. One particular aspect of radiation is that the reflectivity of the surface plays an important role on the amount of heat transferred. Due to differences in the reflectivity of materials and varying pattern densities, different locations will absorb different amount of heat, causing variations in the annealing temperature [13]. The work in [14] investigated the impact of RTA on intra-die variations in the performance, sub-threshold leakage, and other parameters, and demonstrated that most of the observed variations, including inverter delay changes, can be accounted for through RTA-driven variations in the FET extrinsic resistance, R_{EXT} , and V_{TH}. The work in [15] performed thermal simulation and used compact models to analyze within-die variability due to pattern-dependent RTA process. Their simulation showed about 20% variation in the leakage and 3% variation in the frequency due to the non-uniformity of the pattern density in the layout.

Since Ion implantation and thermal annealing are ways to introduce dopant ions into the semiconductor material. This results in variations, both in number and placement of dopant atoms in the channel. Due to the scaling in transistor dimensions, the total number of dopant atoms required to be in the channel to achieve a certain level of doping concentration decreases from generation to generation. As a result the number of dopant atoms required is in the tens or low hundreds for the 45 and 32nm technologies. Therefore,

the variations in the number of dopants around a certain mean value increase significantly. Since ion implantation and thermal annealing are the process steps which affect the number and the distribution of dopant atoms the most, we collectively call this problem random dopant fluctuation (RDF).

2.3.4 Stress

Strain technology, which employs mechanical stress to alter band structure of silicon and reduces carrier effective mass and scattering rate, is introduced to elevate carrier mobility. Because mobility is a strong function of stress, by applying a physical stress on silicon lattice, we can increase the carrier mobility. This increase can lead to a higher saturation current and a higher switching speed for circuits. A tensile stress is desired for NMOS transistors to increase the mobility of electrons, and a compressive stress is desired for PMOS transistors to increase the mobility of holes. Based on the lattice mismatch between Si and SiGe, the bi-axial stress is exerted by depositing a pseudo-morphic Si layer on a relaxed SiGe substrate. On the other hand, the uni-axial stress is applied to one direction, usually to the direction of channel, and has been adopted as standard process since 90nm node because of lower integration complexity and smaller threshold voltage (V_t) shift. The major techniques to introduce uni-axial stress include (i) Embedded SiGe (eSiGe) technology, (ii) Dual Stress Liner (DSL), (iii) Stress Memorization Technique (SMT), and (iv) the parasitic stress from Shallow Trench Isolation (STI). However, stress can also be introduced to the silicon lattice unintentionally. The mismatch in thermal expansion of different materials is one mechanism that can create unintentional stress. The use of shallow trench isolation (STI) is one example. During the oxidation step in the formation of STI, because of volume expansion, the neighboring transistors experience compressive stress. Compressive stress has a negative impact on the performance of NMOS transistors since it greatly decreases the electron mobility. The strain-induced variability is also highly systematic since it depends on the layout of the transistor and its surrounding geometry. The size of the active area and the distance from the gate to the STI edge are especially important when dealing with stress. As the gate moves farther away from the STI edge, it will experience less compressive stress from the expansion of the dielectric material. Larger transistors also

tend to be less sensitive to external stress. As the distance between transistors continues to decrease, the channel gets closer to the STI edge; therefore, a significant increase in unintentional stress on the channel is expected in future technologies. Figure 2.9 shows the effects of layout density on the electron and hole mobility for two 45nm inverters [16]. In some locations, the mobility is degraded by up to 40% (corresponding to the white color and mobility multiplication factor of 0.6) relative to the stress - free transistor, whereas in other locations the mobility is enhanced by up to +20% (corresponding to the black color and multiplication factor of 1.2).

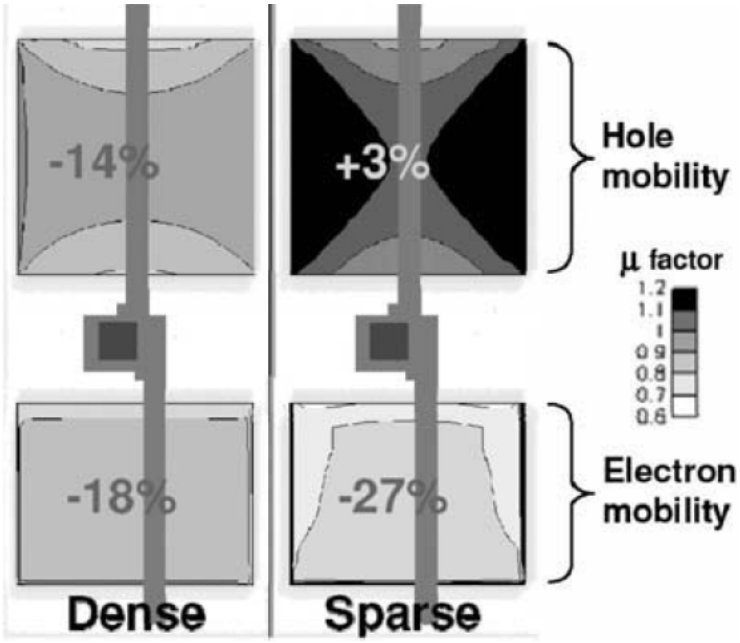


Figure 2.9: Electron and Hole mobility change in different layout environments for 45nm CMOS inverters [16] The color map represents the mobility multiplication factor μ, where darker colors show stress - enhanced mobility, and lighter colors show stress - degraded mobility.

Stress will play more role in degrading the parametric yield in double patterning technology [17]. In double patterning lithography (DPL), overlay error between patterns of the same layer from different exposures translates into line-width/spacing variations as shown

in Figure 2.10 with serious implications on devices and wires. On the device side, the main consequence of overlay error in DPL is its impact on stress. Overlay-induced layout variations that affect stress include:

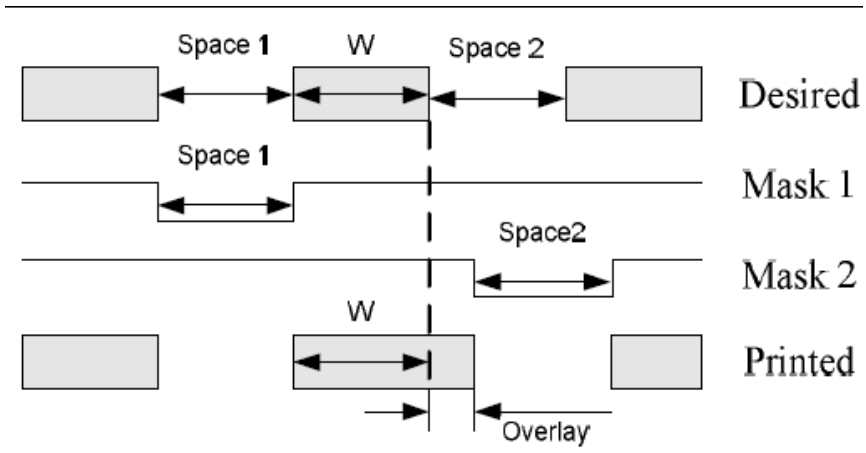


Figure 2.10: Impact of overlay error on CD uniformity

- Gate spacing affecting mechanical stress from stress liner.
- Gate-to-contact spacing with impact on source/drain resistance, gate-to-contact capacitance.
- Shallow trench isolation (STI) width, which impacts STI stress.
- S/D length influencing embedded SiGe and STI stress sources

2.3.5 Chemical-Mechanical Polishing

All of the variations sources described so far are part of the front-end process, which refers to the steps that create the actual transistors. Chemical-mechanical polishing (CMP) is used repeatedly in the back-end process, which refers to the steps that form the wiring and interconnect of the circuit. CMP is used to achieve smooth and planar surfaces from which subsequent layers are able to be fabricated. Decreasing depth-of-focus in modern

lithography systems underscores the need for exquisite planarity and without such planarity, features to be patterned may be out of focus due to surface height fluctuations (nanotopography). However, CMP is not a variation-free process itself: it is a significant source of systematic variations resulting from both process conditions, including variations in down force, rotational speed, pad conditioning, and temperature as well as designed feature sizes and pattern dependencies. Two kinds of variations are most common in the CMP process: dishing of copper and erosion of dielectric [18] both are shown in Figure 2.11.

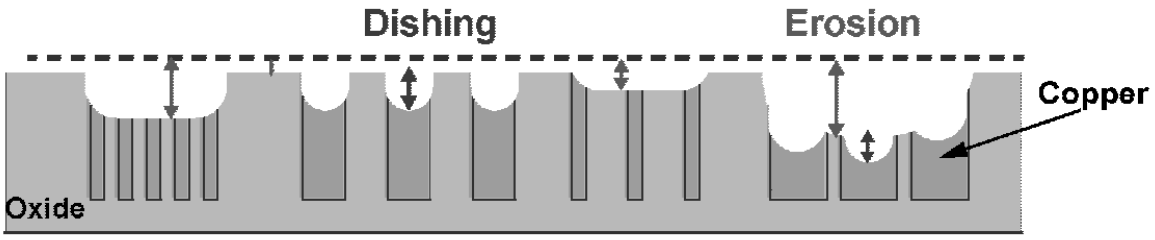


Figure 2.11: Dishing and erosion in CMP

Dishing refers the over polishing of the features within the trench relative to the surface of the dielectric layer. Erosion refers to the removal of surrounding dielectric when it should not be removed. In general, larger features suffer more dishing than smaller features, but conversely, smaller features suffers more erosion compared to the larger features. For medium-sized-features, both dishing and erosion contribute to some degree of polishing variations. CMP-induced variability is highly systematic, since it relates directly to the feature size and layout pattern densities being polished.

2.4 Modeling Process Variations

Thus, we find that the advanced CMOS process technologies introduce performance variability, which causes severe variability in the performance of advanced VLSI circuits and systems. Therefore, it is critical to accurately model process variability when predicting the

performance of advanced VLSI circuits. Furthermore, accurate compact MOSFET models to account for process variability in VLSI circuits are indispensable. A compact variability model must accurately describe the process-induced device and circuit performance variability, and it must be physics-based and predictive. For a robust design and high yield, it is essential for the process-design kit (PDK) to support the worst-case fixed and user-defined corner models, statistical models, and Monte Carlo models for yield estimation of a design for a target technology.

2.4.1 Worst-case corner models

The worst-case corner models are generated by offsetting the selected process-sensitive compact model parameters by a fixed number, n, of the standard deviation σ for each parameter, to account for the window of process variability. For example, the corner models include $V_t = V_{t0} + n\sigma$, where V_{t0} is a selected model parameter of the typical (TT) model, and n is selected to set the fixed lower and upper limits, LL and UL, of the worst-case models. Typically, the TT model is generated from the measured data on a single golden wafer of the center-line process. Thus, the worst-case corner models give designers the capability to simulate the pass/fail results of a typical design and are usually pessimistic. Conventionally, process variability is modeled on the basis of the worst-case four corners – two for analog applications and two for digital. The corners for analog applications are generated from slow NMOS and slow PMOS (SS) to model the worst-case speed, and from fast NMOS and fast PMOS (FF) to model the worst-case power. The corners for digital applications are generated from fast NMOS and slow PMOS (FS) to model the worst-case 1, and from slow NMOS and fast PMOS (SF) to model the worst-case 0.

In this modeling approach, the standard deviation limits are preset pessimistically to include any potential process variability over a wide range. Most foundries support worst-case models, in which the Spice parameters are set to the specific LL and UL of the corresponding process parameters – for example, BSIM4 V_{TH0} corresponding to $V_{TH,N}$ and $V_{TH,P}$ for NMOS and PMOS, and LINT corresponding to L (where LINT is the channel length- modulation Spice model parameter that defines L_{eff}). The process parameters are specified in the design documents and are based on a 3σ or 6σ parameter distribution.

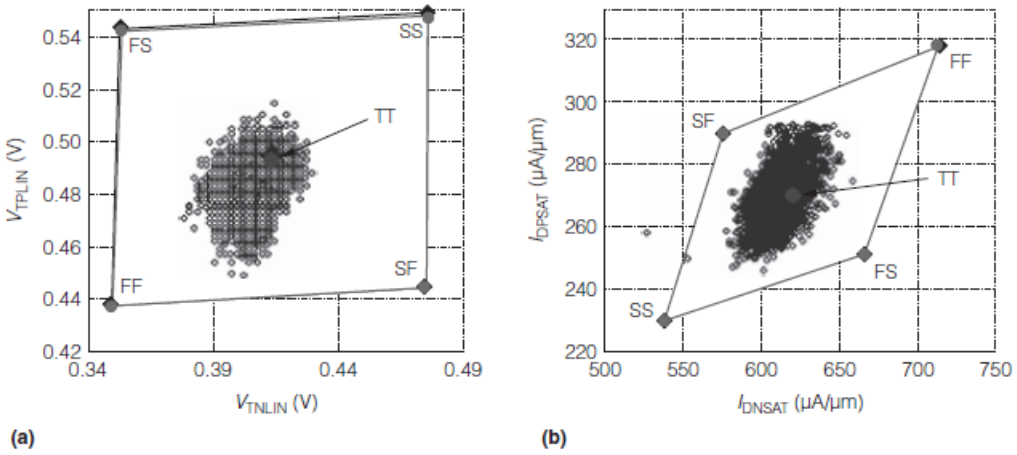


Figure 2.12: Production data distribution and simulation data generated using fixed-corner models: NMOS linear V_{TH0} versus PMOS linear V_{TH0} (a) and NMOS saturation current (I_{DNSAT}) versus PMOS saturation current (I_{DPSAT}) (b) [19].

Figure 2.12 shows plots obtained for typical industry standard fixed-corner models [19].

As is evident from the figure, the fixed-corner method is too wide, so it could end up rejecting a valid design. The major problems with the worst-case corner models, on the other hand, are that in most cases the existing correlations between the device parameters are ignored and that the models include pessimistic corner values. This over-pessimism makes the design problem more difficult to solve.

2.4.2 Statistical corner models

Statistical corner models are generated using data from different dies, wafers, and wafer lots collected over a long enough period of time to represent realistic process variability of the target technology. The corner parameters are obtained by adding a realistic σ of the corresponding model parameter to its TT value, where the value of each σ is obtained from the distribution of a large set of production data. Thus, statistical corner models help designers perform a realistic pass/ fail evaluation of a design. The electrical data is collected from multiple devices, wafers, and lots over a long period of time. The collected data may

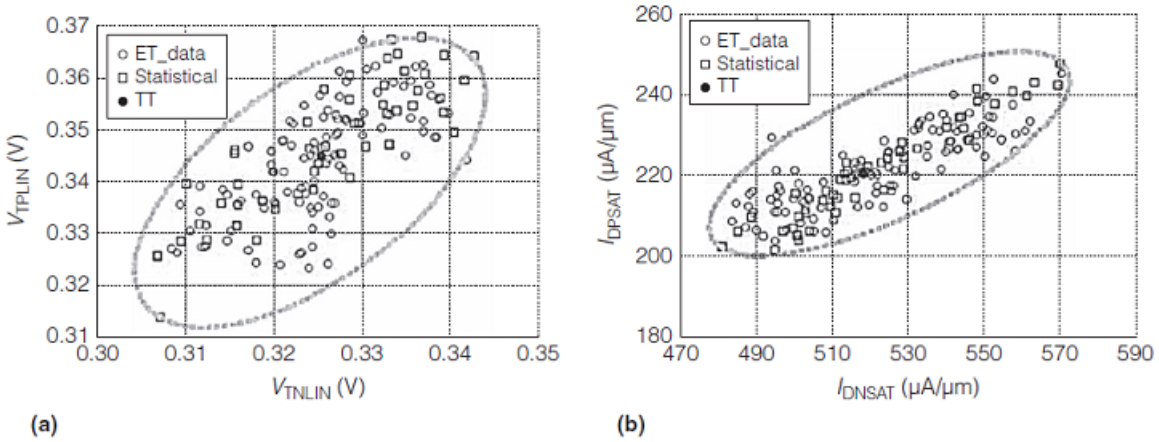


Figure 2.13: Production data distribution and simulation data generated using statistical models: NMOS linear V_{TH0} versus PMOS linear V_{TH0} (a) and NMOS saturation current (I_{DNSAT}) versus PMOS saturation current (I_{DPSAT}) (b) [19].

be current-voltage and capacitance-voltage characteristics or electrical test (ET) data for process monitoring. To use current-voltage and capacitance-voltage data for modeling, the physical effect causing process variability must first be determined. Then the model parameters are extracted from the devices with the measured data at the distribution boundaries. However, this approach is time consuming, so ET data is normally used to generate an efficient statistical model. In the case of a new technology generation, the production data for statistical modeling is limited; therefore, systematic technology CAD (TCAD)-based process variability data can be generated for statistical corner modeling. Statistical models represent realistic process variability and can pass a valid design rejected by pessimistic fixed-corner models.

Figure 2.13 shows the distribution of production data of an advanced CMOS technology, along with simulation data obtained via statistical models generated for the same technology. To further improve the robustness of the statistical models, an appropriate percentage of guard-banding can be used to address any random variability over some specified period of time. Multiple corner statistical models can be generated from a set of wafers, each representing production data at the outermost boundary of the distribution

in the database. Taking the statistical variations of physical parameters and modeling the effect on the delay of standard cells was shown in [20]. On the other hand, statistical timing analysis is similar to deterministic (static) timing analysis in that arrival times are propagated through the circuit from primary inputs to primary output. In statistical timing analysis, however, the gate delays and arrival times are represented with random variables [21]. The difficulty of statistical timing analysis results from the correlations that arise among the arrival times in the circuit and between the arrival times and gate delays. These correlations must be taken into account when arrival times are propagated in the circuit, leading to an exponential run time complexity and making statistical timing analysis a challenging problem.

Monte Carlo SPICE models allow for directly estimating the yield of a given VLSI design. Such models include both a local and a global parametric distribution. The local distribution accounts for the intra-die process variability or mismatches of each selected model parameter, and it models device-device mismatches by calculating the variance from the random distribution within the target limits, such as W = W ± CD and L = L ± CD. The global distribution accounts for the inter-die or lot-to-lot process variability of each selected Spice model parameter. In Monte Carlo simulations, critical device parameters are randomly distributed according to process specifications such as VTH and Leff. The basic parameters are allowed to vary independently and include device correlation. These simulations make it possible to simultaneously test a design for process variability and mismatches.

2.4.3 Systematic Variation-Aware Modeling

If the physical source of variability varies within a die because the die is large relative to the wafer, or because of a strong layout dependence then the task of determining the design performance variations becomes more difficult because the number of entities varying is larger and simple worst case analysis is not possible [22]. The authors of [23] provided two approaches to create variability-aware timing models for standard cells. For both approaches it is assumed that all transistors within a cell experience identical process effects. The first approach utilizes geometrical biasing of transistor L and W in standard

cells to create delay sensitivity tables. The second approach combines rigorous process simulation with contour based timing characterization to develop compact parameter delay models for the cells. In [24] context-aware timing analysis is proposed by pre-characterizing the library cells where 81 versions of each cell is included in the cell library to account for different contexts. This method overcomes the magnitude of the pessimism of traditional static timing analysis which neglects systematic components of ACLV. This can amount to as much as 40% tightening of the best-case to worst-case timing spread.

2.5 Impact of Process Variations on Functional and Parametric Yield

Variability affects IC yield. Yield is defined as the probability that a chip is both functional and meets the parametric constraints, such as timing and power. In principal, a circuit with more design margins will have a higher yield. The challenge is in finding the smallest margin necessary for the required yield so that performance is not overly constrained.

Process variations affect both functional and parametric yield. *Functional yield* (also known as hard or catastrophic yield) is when the process variations destroy the functionality of a circuit. The variations might cause short circuits, open circuits, or other types of binary failures. Examples of functional yield loss mechanisms are defect particles, lithography hotspots, and CMP hotspots. Numerous work on the literature tools that try to solve these problems. [25–28]. On the other hand, the natural variability of the process, as well as the non-catastrophic impact of some types of defects, will lead to a spread of the various device parameters, and this spread will in turn result in a spread of IC performances. Performance is usually quantified by metrics such as speed of execution or power consumption. *Parametric yield* measures the ratio of chips with performance that meets the product performance constraints. When studying the variations in transistors and interconnects one should focus on: Vt, Tox, μ, device L and W, and interconnect R and C. The implied targets are Ion and Ioff and the corresponding delay and power dissipation associated with them.

In the remaining of this work, when handling parametric yield we will focus on transis-

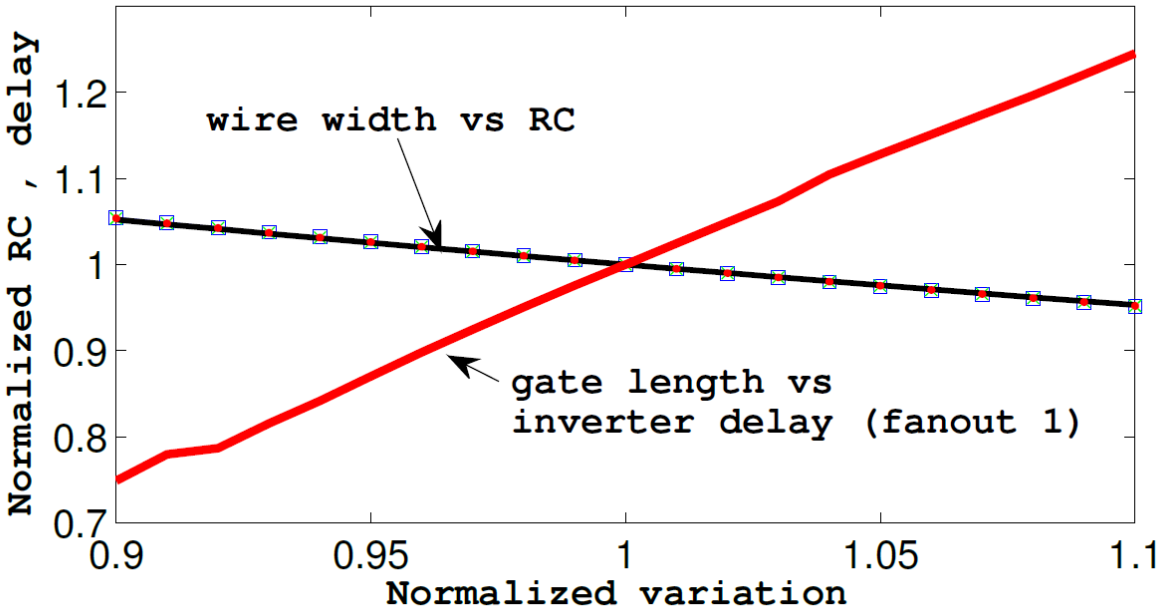


Figure 2.14: Comparison of the sensitivity of delay to the variations in interconnect wires width and transistor gate length [29].

tor's variations and ignore the variation sources that affect the interconnects. We do that because it has been shown that the impact of gate length variations is 5X the impact of wire width as shown in Figure 2.14 [29]. It was also verified that power are not sensitive to line-width variation of the wires. It is worth noting that the impact of wire width variation is expected to be even less because of the cancellation of power and delay variation due to averaging over long wires.

To illustrate the impact of variations on actual products, Figure 2.15 plots the normalized distributions of frequency and standby leakage of Intel microprocessors on a single wafer [30]. Parameter variations result in greater than 30% frequency spread and 5X variations in chip leakage at 130nm technology. For a production 65nm IBM microprocessor the variations increased to 50% frequency spread and 10X variations in leakage power as

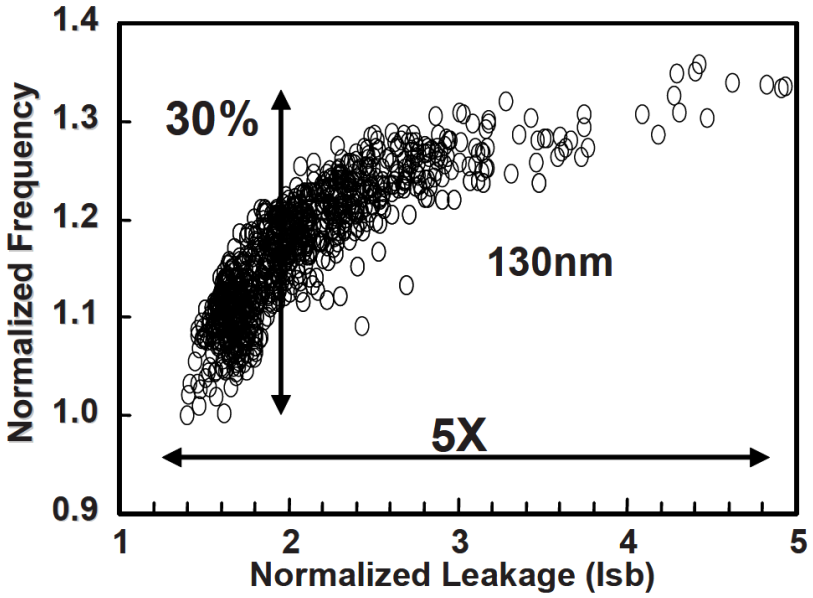


Figure 2.15: Frequency and leakage variations of a 130 nm microprocessor [30]

shown in Figure 2.16 [31]. The large frequency spread necessitates expensive frequency binning in which each chip is tested to determine its maximum frequency and power before it can be sold. This binning is an expensive and time-consuming process. As a result, yield is affected by parameter variations: chips that operate too slowly with high standby leakage power, or those that have high performance but are above the power envelope, must be discarded. Microprocessors often represent extreme examples of semiconductor engineering. The problem is more generally valid; performance and power are significantly impacted by unmitigated parameter variations resulting in parametric yield loss.

Where yield is quality dependent metric, *Reliability* is a time dependent metric. Reliability is a characteristic of a product that is associated with the probability that it will perform its intended function under specified conditions for a stated period of time. Although the reliability considerations across the electronics industry is multi-faceted and is a vast topic, in general they can be classified in three groups [32]. The first group involves permanent damage arising from generation of bulk defects in SiO2 that leads to gate dielectric breakdown in logic transistors, anomalous charge loss in Flash transis-

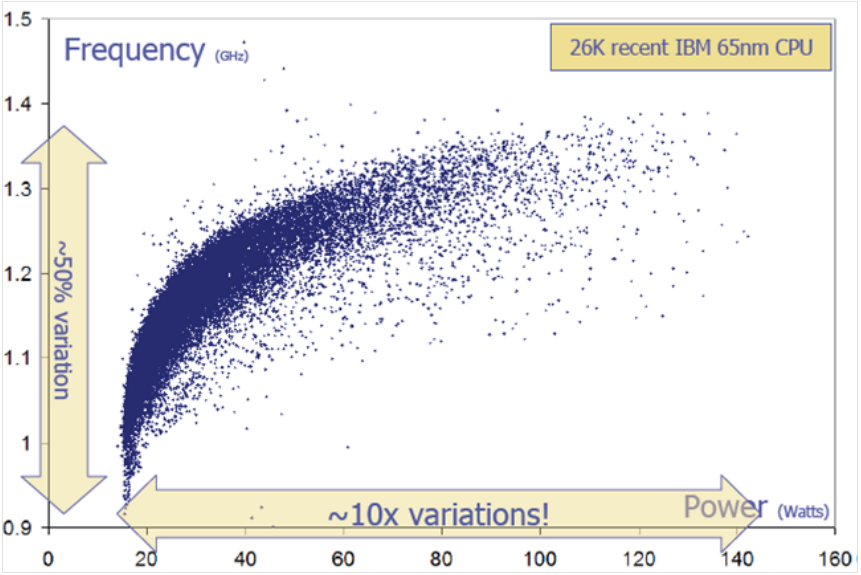


Figure 2.16: Frequency and leakage variations of a 65nm microprocessor [31]

tors, loss of resistance ratio in MRAM cells radiation induced permanent damage in SRAM cells. The second group involves permanent damage arising from loss of passivated surfaces (e.g., broken Si-H bonds for negative-bias temperature instability or hot carriers injection (NBTI/HCI) damages in microelectronic and macro-electronic applications loss of Si-H bonds and increase in dark current in amorphous-Si based solar cells, etc.). There is a third group of reliability issues involving transient errors (e.g., soft error due to radiation). The yield and reliability of microelectronics manufacturing products are highly related, but high manufacturing yield due to one manufacturing process does not necessarily imply high reliability of the products from that manufacturing process in the field [33]. Since, process variations negatively impact both yield and reliability, in our study we will aim to mitigate of variations on both.

2.5.1 Process Variations Effect on Transistor Characteristics

The equations (2.2) - (2.4) below highlight some of the most important benchmarks in determining transistor performance [34]. Equation (2.2) describes the saturation current

I_{sat} which can be used to evaluate the drive strength of transistors. Equation (2.3) describes the leakage current I_{leak}, which can be used to evaluate the leakage power consumption during the idle stages. Equation (2.4) is the delay equation, where the delay constant τ is used in the ITRS roadmap to characterize transistor switching speed.

$$I_{sat} = \frac{1}{2}\frac{W}{L}\mu C_{ox}\left(V_{gs} - V_t\right)^2\left(1 + \lambda V_{DS}\right) \tag{2.2}$$

$$I_{leak} = \mu\frac{W}{L}\left(\frac{kT}{q}\right)^2 e^{\frac{-V_t}{\frac{mkT}{q}}}\left(1 - e^{\frac{-V_{ds}}{\frac{kT}{q}}}\right) \tag{2.3}$$

$$\tau = \frac{C_{load}V_{supply}}{\frac{1}{2}\frac{W}{L}\mu C_{ox}(V_{gs} - V_t)^2} \tag{2.4}$$

where W is the transistor width, L is the transistor gate length, μ is the mobility, V_t is the threshold voltage, C_{ox} is the gate oxide capacitance, λ is the channel length modulation parameter and kT/q is the thermal voltage.

Here, we can clearly see how each transistor parameter W, L, C_{ox} and V_t affects the performance of the transistor. This will help us to relate the variations in transistor parameters directly to these performance metrics. In the last decade the 3σ variations of the effective channel length (L_{eff}) increased from 32% to 47%, the 3σ variations of the gate oxide thickness (T_{ox}) increased from 8% to 16 %, and the 3σ variations of threshold voltage (V_t) increased from 5% to 16% [35]. The variations are defined as the ratio of 3σ to the nominal value. It is evident that the variability is increasing with technology generations. Furthermore, it can be seen that while T_{ox} and V_t observe moderate variations increase, Leff experiences large variations increase. The within-die portion of the process variations are also increasing, for instance, with technology scaling, the channel length variations caused by the within-die variations raised from 40% to 65%. Tables 2.1 and 2.2 shows the increasing trend of variability with the advanced technology nodes.

Table 2.1: Intra-die V_t variability increase with technology node

Parameter						
Leff (nominal)	250 nm	180 nm	130 nm	90 nm	65 nm	45 nm
V_t (nominal)	0.450 V	0.400 V	0.330 V	0.300 V	0.280 V	0.200 V
V_t (σ)	21 mV	23 mV	27 mV	28 mV	30 mV	32mV
V_t (variations)	4.7 %	5.8 %	8.2 %	9.3 %	10.7 %	16 %

Table 2.2: CD variability budget with technology node

Leff (nominal)	65 nm	45 nm	32 nm	22 nm
Total gate (3σ)	2.5 nm	1.9 nm	1.3 nm	0.9 nm
Lithography (3σ)	2.2 nm	1.4 nm	1.1 nm	0.8 nm
LER (3σ)	2.0 nm	1.4 nm	1.0 nm	0.7 nm
Gate Etch (3σ)	1.1 nm	0.8 nm	0.6 nm	0.4 nm

2.5.2 Process Variations Impact on Circuit Timing and Leakage Power

In the previous section we reviewed how the variations in the transistor parameters would impact the transistor performance. In this section we will see how these variations will impact the circuit performance. Process parameter variations cause a large spread in the timing and leakage distributions of circuits [30, 31]. Table 2.3 breaks down the overall variations into contributions from individual parameters. Variations in V_t and channel length contribute most heavily to overall variations [36]. Authors of [14] reported up to 20% variations in 65nm inverter delay they explained these variations by variations in the annealing temperature during RTA because of different layout densities.

Moreover, different designs react differently to process variations according to their function, their design style and their constraints [36]. Also Alioto et al. in [37] found that Domino logic circuits (though faster compared to static logic) suffer from a 2X higher variability compared to static CMOS logic. This can be explained because of the positive feedback effect the keeper transistor in the Domino logic tends to amplify the variations.

Table 2.3: Average contributions of variations from individual parameters over various circuits. [36]

Parameter	Delay $(\frac{\sigma}{\mu})\%$	Power $(\frac{\sigma}{\mu})\%$
t_{ox}	1-2%	1-2%
W	$<<$ 1%	0.5-1%
L	3%	$<$ 2%
V_t	2.5-6%	1.75-4.75%

These variations have a huge impact on the yield of the dies. Dies with a high delay and high leakage power consumption must be discarded. Dies with an acceptable standby leakage are binned based on their frequencies and are priced accordingly. The large variations in the standby leakage current are mainly due to variations in the sub-threshold leakage, which is the main contributor to the leakage current. Because of the inverse exponential relationship between the sub-threshold leakage and the threshold voltage, small variations in V_t results in large variations in the leakage current. In the high-performance CMOS design, the leakage power consumption can be responsible for 40% or more of the total power consumption of the circuit.

2.6 Conclusion

In this chapter, the different sources of process variations are reviewed. It is shown that with continued scaling lithography induced intra-die variations are growing in significance to the point where they are dominant. We also showed that the variability is getting so critical that it must be taken into account at all stages of the design. Intra-die variations are both random and systematic. We have also shown how systematic variations are becoming the source of variation in newer technologies. It is important to determine the systematic part of variability which can be mitigated to a certain degree most of the time. We also showed that transistor's variations are more critical than the effect the variations in the interconnects. Lithography is the biggest challenge in advanced nodes and is the most contributing process effect to the systematic variation. For these reasons

we are going to focus our work on the analysis of intra-die systematic variations and in the remaining chapters we will study how to detect, mitigate and fix problems resulting from lithography. In the next chapter we will review the state-of-the art DFM techniques and how they are introduced in different stages of the design to mitigate and fix the effects of process variability.

Chapter 3

State-of-the-art Variation Mitigation Techniques

3.1 DFM evolution

Design for manufacturability (DFM) in its broad definition stands for the methodology of ensuring that a product can be manufactured repeatedly, consistently, reliably, and cost effectively. This is achieved by taking all the measures needed for that goal starting at the concept stage of a design and implementing these measures throughout the design, manufacturing, and assembly processes. Back until the 0.18μm generation, the interface between the design and manufacturing phases of an integrated circuit was well represented by straightforward device models and simple geometric design rules which typically determined the minimum widths and spacings for the various layers that composed the integrated circuit. The role of the models was to enable us to predict the behavior of the integrated circuit given that it is not possible to prototype the IC in order to find out whether and how well it works. The role of the design rules was to insure that the yield of the circuit - defined as the proportion of manufactured circuits that are functional and meet their performance requirements - was economically viable. The relationship between yield and design rules existed because the yield loss mechanism in those manufacturing processes was dominated by topology changes (shorts and opens) caused by particulate

contamination and similar phenomena. As scaling continued, our ability to reliably predict the outcome of a semiconductor manufacturing process has steadily deteriorated and process complexity and the challenges of accurately modeling variability have conspired to increase the error in performance predictions, leading to a gap in model-to-hardware matching.

With older technology, Boolean-based design rules checks (DRC) worked well and have been the design sign-off to guarantee a manufacturable design. As technology scales, problems arise. Pattern distortion due to the optical proximity effect becomes more pronounced as the technology goes deeper in the sub-wavelength regime. For the 90nm node, the first-order proximity effect is to the structures immediately adjoint to the polygon of concern. At the 45nm node, the proximity effect influence is as far as a few structures away from the polygon. When the proximity effects are extending far, it is very difficult to code Boolean-based rules to describe this effect so that designers can design for it. Starting at 45/40nm, the increasing complexity of DRC and DFM rules began to stress traditional physical design flows. This trend is expected to continue and worsen at the 32nm and 22nm nodes, where manufacturing closure may become a serious bottleneck in design schedules. To ensure that layouts are made lithography-compliant while maintaining design intent, lithographers have been working more closely with layout engineers to eliminate non-RET-compliant layout patterns from the design. Through mutual collaboration, they have defined a set of nonsimple DFM rules that extend the application of design rules to create RET-compliant ASIC designs.

This trend of adding more complex design rules in response to the non-monotonic layout sensitivities experienced in low k1 lithography processes has led to the escalation in design rule complexity. The forbidden pitches introduced by using off-axial illumination (OAI) are avoided in layout designs by introducing a rather complex set of multifeature width-dependent spacing rules. As a result, what used to be a simple pass-fail limit is now a complex problem. Unfortunately, even with this added complexity, design rules cannot provide absolute assurance that a design-rule clean layout will yield or perform adequately. Unanticipated asymmetric width-space combinations or 2-D constructs not considered during the design rule definition process, have led to yield losses when seen in design-rule-clean layouts. Meanwhile, efforts to simplify the design rules will eliminate

some layout constructs that are extremely valuable to a particular design even though they might be adequately manufacturable.

In essence, the rule-based design has to make a compromise of either being too conservative or being too complex. Also DFM techniques as via-doubling and dummy fill are now standard in 45nm technologies to overcome low via yield and CMP problems respectively. Recently, layout analysis methods, simulation tools, and corresponding models are being developed to extensively analyze layouts in order to predict the locations of yield detractors that are more popularly termed hotspots. These techniques include critical area analysis which determines the sensitivity of layout patterns to random spot defects, post-OPC through-process printability verification to find lithography and etch hotspots, density checks to find chemical mechanical polishing (CMP) hotspots, and so on.

In the 65-nm and 45-nm nodes, particularly for high-performance process flavors, silicon providers are providing variant guard-bands at the level of device (SPICE) model or interconnect RCX models, corresponding to different regimes of manufacturing-friendliness or DFM score in the tape-out. A first example might be the reduction of worst case - best case (WC-BC) guardband for RC extraction, which is enabled by the deployment of new golden models for chemical-mechanical planarization (CMP), which lead to new process-aware extraction and timing analysis (as well as process-driven dummy fill) flows. A second example might be the application of a different (narrower) SPICE model guardband for, e.g., a multifingered device that is laid out with optimal (restricted) pitch and poly dummy layout choices.

3.2 DFM vs Statistical design

Design-oriented strategies can be classified into design for manufacturability (DFM) approach and Statistical design approach. This first design strategy takes advantage of the parts of design that are model-able (or in other words, systematic). For example, it is well-known that the transistor orientation impacts the fabricated channel length of the transistors. So in the case of analog circuits where matching between transistors is important, designers will not use transistors with different orientations as a simple DFM proactive

measure. The ultimate goal for engineers is to be able to approach all of the problems using DFM solutions. This requires the understanding and investigation of variation sources and ultimately the incorporation of these findings into modeling.

For the type of variations for which the source is either unknown or is truly random, we can use a second design approach called statistical design. This design approach follows the principle of better-than-the-worst-case-design. In the past, designing for the worst-case was common. For example, in a digital integrated circuit, in order to achieve a high yield, designers are forced to put large margins into their designs to ensure that the slowest logic path can still operate under the frequency constraint. However, it becomes exponentially more expensive to accommodate the slower tail of the distribution. The key concept of statistical design is not to lose too much performance accommodating a small percentage of circuits, but rather to make engineering trade offs between performance and statistical yield. Post-fabrication testing is necessary if a statistical design approach is used during the design process.

3.3 DFM Techniques

Mitigation of process variation is best divided into pure process techniques (i.e., techniques transparent to design), process-design co-optimization techniques (i.e., techniques that exercise tight cooperation between process and design), and pure design techniques (i.e., techniques transparent to process). Targeting the process itself involves altering process modules and/or flows of device design, directly impacting variation at or close to the source. Examples of pure process mitigation techniques include targeting key transistor properties to reduce random dopant fluctuation, reducing traps at the HiK+MG interface to reduce random charge variation, improving patterning techniques to reduce LER and end-cap variation, and improving polishing technologies to reduce systematic cross-wafer variation [38]. Examples of combination design-process techniques include optimizing topology, using optical proximity correction to reduce random and systematic variation, and adding dummy features to reduce systematic variation. Pure design techniques include chopping and auto-zeroing to compensate for random variation and common-centroid layout to compensate

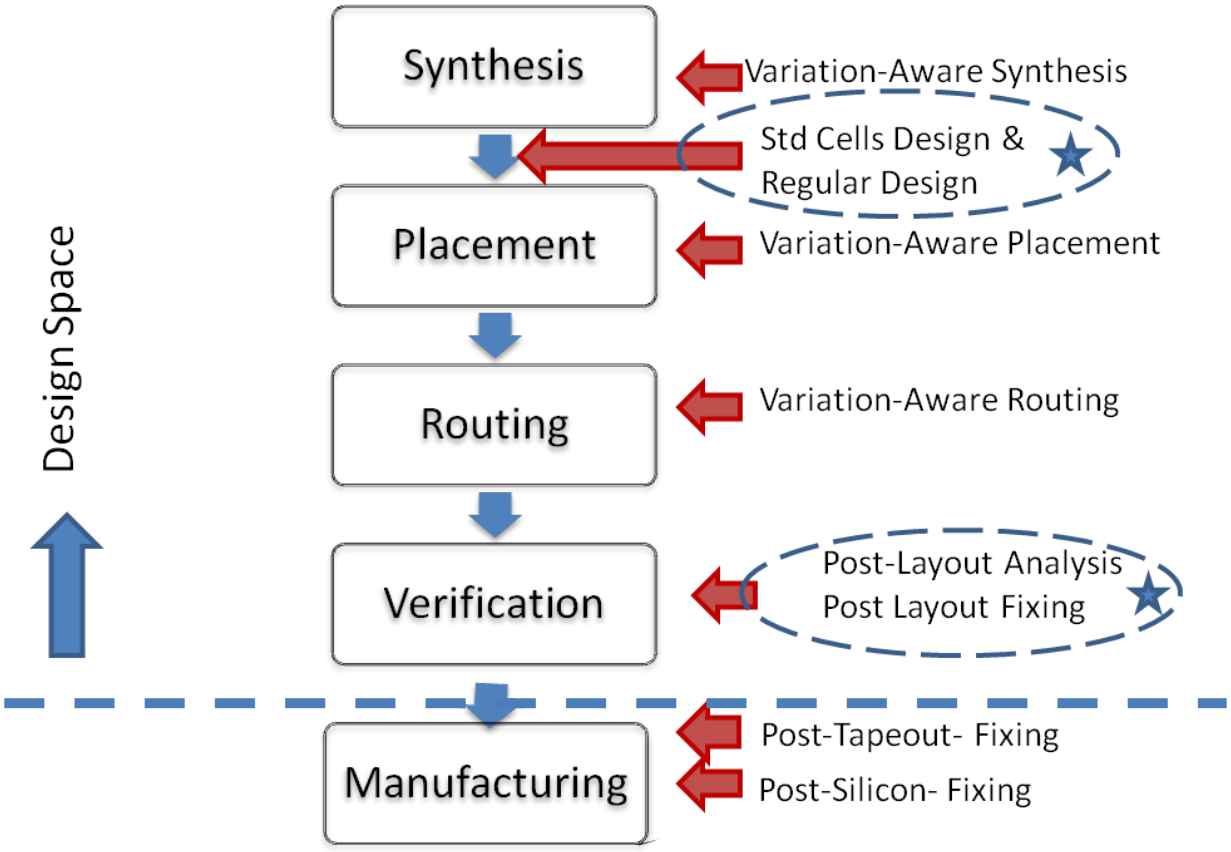


Figure 3.1: DFM techniques at different design stages.

for systematic variation [38].

Prediction and compensation of systematic variation have traditionally been done by the manufacturing process, with only simple guard-banded abstractions (e.g., design rules) being passed on to the designers. However, the increasing magnitude and 2-D pattern dependence of these variations, their impact on design metrics, and the inability of manufacturing equipment and process techniques to fully mitigate them, are causing serious concern in sub-100-nm technologies. A number of recent works have proposed systematic process variation-aware design and analysis to close the loop from manufacturing simulation back to the design flow. These DFM techniques targeting mitigation of variation effects on the yield of the circuits are introduced along the various stages of the design as

shown in Figure 3.1. Below we will review recently published work classified according to the stage of design they are addressing with special focus on the areas we will be addressing in the next chapters.

3.3.1 Variation-Aware Synthesis

In the domain of high-level-synthesis, process-variation-aware research is still in its infancy. It is important to raise process variation awareness to a higher level, because the benefits from higher level optimization often far exceed those obtained through lower-level optimization. Furthermore, higher-level analysis enables early design decisions to consider lower-level process variation, avoiding late surprises and possibly expensive design iterations. A proactive methodology for defeating manufacturing problems is proposed in [39], which is not a post process. Nardi et al. in [39] proposed a logic synthesis for manufacturability. This methodology introduces the manufacturability cost into logic synthesis and replaces the traditional area-driven technology mapping with a new manufacturability-driven one. It realizes larger reduction of the manufacturability cost when yield-optimized cells are available in the cell library.

3.3.2 Standard Cells and Regular Design

We will show in next chapter how regularity-enhanced design is friendly to photolithography. However, design restriction associated with regularity reduces design flexibility and requires extra features such as dummy patterns. From a viewpoint of the circuit performance, regularity-enhanced design has negative impact, since extra features require extra area and dummy patterns increase parasitic capacitance. Therefore designers should consider the trade-off between the advantage in printability and the disadvantage in circuit performance. Sunagawa et al. [40] discuss the effect of regularity on designs of 90nm, 65nm and 45nm. A regularity-enhanced standard cell with dummy poly insertion does reduce performance variability. However, the amount of the improvement is moderate, and, more importantly, noticeable amount of performance overhead is observed, which means that

this level of regularity does not pay off in the 90nm process. In a 65nm process, layout regularity also helps to suppress performance variability while it incurs performance penalty of 4% speed loss in a ring oscillator circuit estimated by lithography and circuit simulations. In a 45nm process, on the other hand, certain level of regularity is indispensable for ensuring printability under adequate amount of lithographic process windows.

From those experimental results, as the technology scaling progresses, the required level of regularity becomes ramping up steeply. It is important to evaluate the minimum amount of layout regularity that is necessary for securing required level of printability. On the other hand, layout systematics are dominated by lithography and stress-related effects that have a spatial range from 200nm-1000nm and 1000nm, respectively. These large interaction ranges make it very difficult to account for this shift in performance at the cell level using conventional design rules as the layout context around the cell is unknown. Instead, a layout methodology that can ensure uniformity across longer ranges is desired. Jhaveri et al. report that there are two broad classes of regularity in the design, namely micro-regularity and macro-regularity [41]. The micro-regularity relates to the number of different layout constructs, such as line-ends, used to implement a given design. Layout constructs are localized layout shapes such as line-ends, L-shapes, T-shapes, and so on. Logic-designs are created using a set of design rules that specify a set of illegal constructs. Anything not explicitly prohibited by the rules is allowed to be used in the layout of the design. The corresponding layouts are not micro-regular. On the other hand, restricted design rules (RDRs) have been gaining momentum in the industry. Poly is the first layer that is becoming micro-regular throughout the industry. At the 65nm technology nodes, transistor gates are required to be unidirectional, whereas at 45 nm and below strict pitch requirements are also being enforced on transistor gates. Application of micro-regularity to other design layers such as metals has also been demonstrated in industry. Intel has demonstrated the use of micro-regularity across all design layers for the 45 nm technology node [38].

Macro-regularity relates to the total number of layout patterns present in the layout. A layout pattern is defined as a set of all layout shapes contained within an optical interaction range of a given layout construct. This optical interaction range varies from 200 for patterns that are formed from deviations to an underlying fabric to 1000 nm for arbi-

trary layouts. An obvious example in the efforts of regularity is the SRAM design. Figure 3.2 shows the evolution in the SRAM cell design to one directional highly regular shapes. The restriction of unidirectional features, uniform gate dimensions and gridded design are shown in the 45nm and 32nm designs in Figure 3.3 [42].

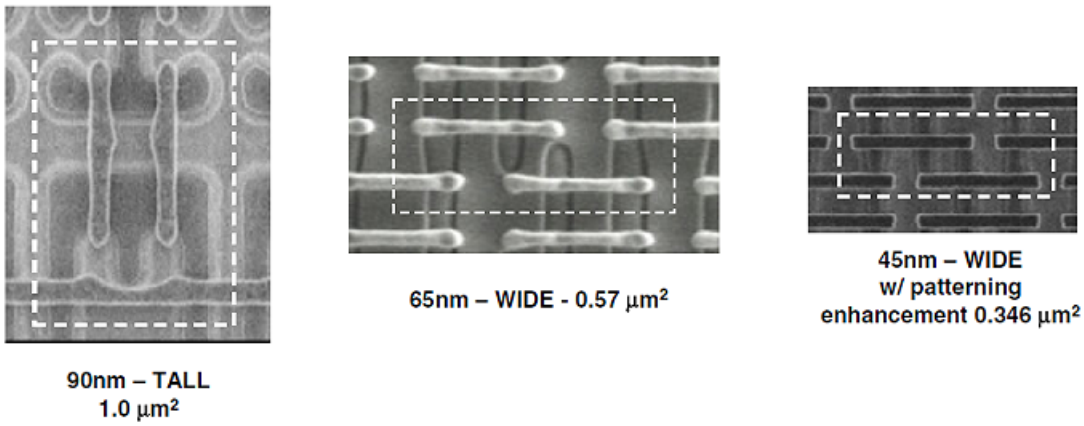


Figure 3.2: SRAM Cell topology [38]. Efforts of regularity are obvious starting from 65nm and below.

Jhaveri et al. in [41] proposed a shift toward regular design fabric. The regular design fabric and templates specifies the allowed layout constructs and layout neighborhoods using a completely prescriptive approach to IC layout design. The fabric specification consists of a set of allowable layout constructs and a design grid, which provides a list of valid locations in the 2-D space for specific constructs to be placed at each layout layer. Since the number of layout constructs is implicitly controlled through the selection of layout shapes and layout grids, the regular design fabric satisfies the micro-regularity constraints. To ensure macro-regularity and enable efficient designs using regular design fabrics, they proposed to add a level of abstraction between the standard cells and the regular design fabric, which is called logic templates. The pre-qualified templates are assembled into the required larger functions of a fully functional cell library, including commonly used standard cells as well as larger logic functions, also known as bricks. The use of the right

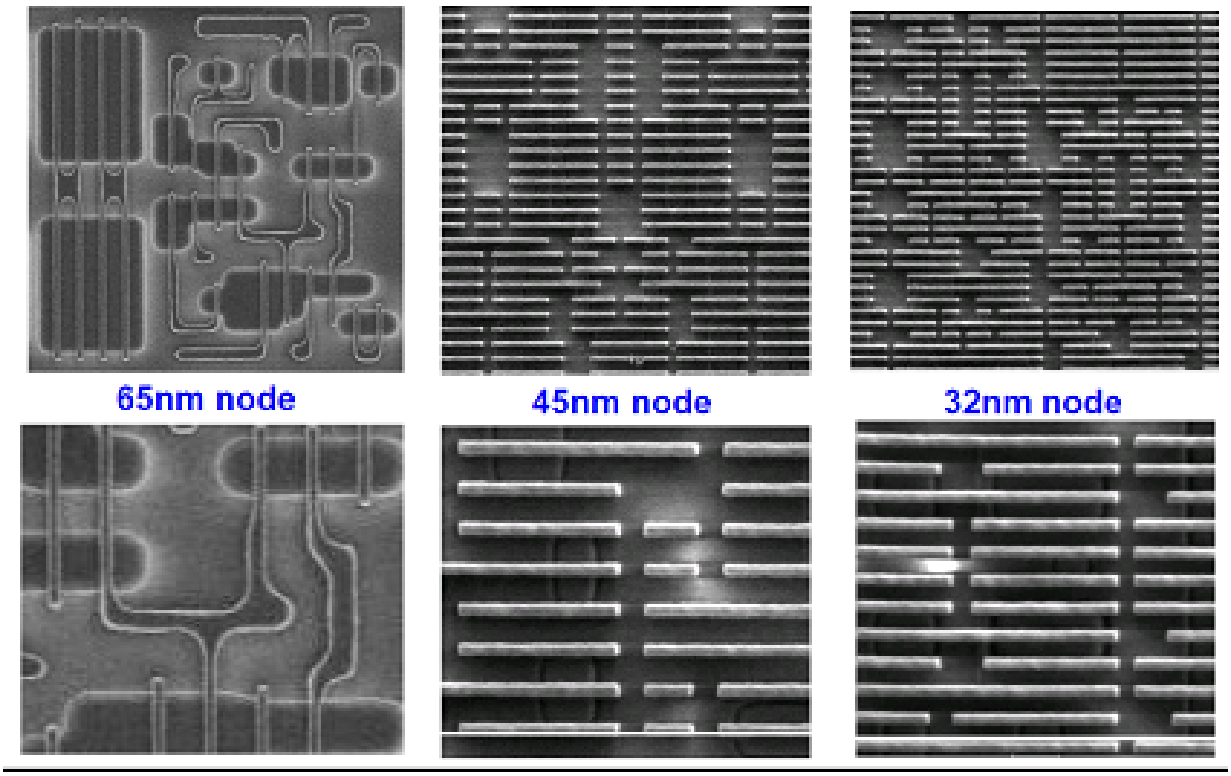


Figure 3.3: Layout restrictions in logic design [42]. The restriction of uni-directional features, uniform gate dimension and gridded layout are shown in 45nm and 32nm designs

set of bricks can enable more efficient mapping of a given logic design as compared to a conventional standard cell library. (see Figure 3.4.)

The standard cell library or a bricks library created using logic templates is used with commercially available synthesis and place-and-route tools to assemble IC designs and so the proposed flow does not need any modification to existing design flows. The real challenge of this flow is defining a fabric and mapping the 70 logic functions to create logic templates. The described methodology does not blindly map layouts on uni-directional uniform grids, but instead relies on close collaboration between circuit designers, process experts, and layout designers to select the right set of patterns that will enable die cost scaling from node to node as well as meet the product requirements.

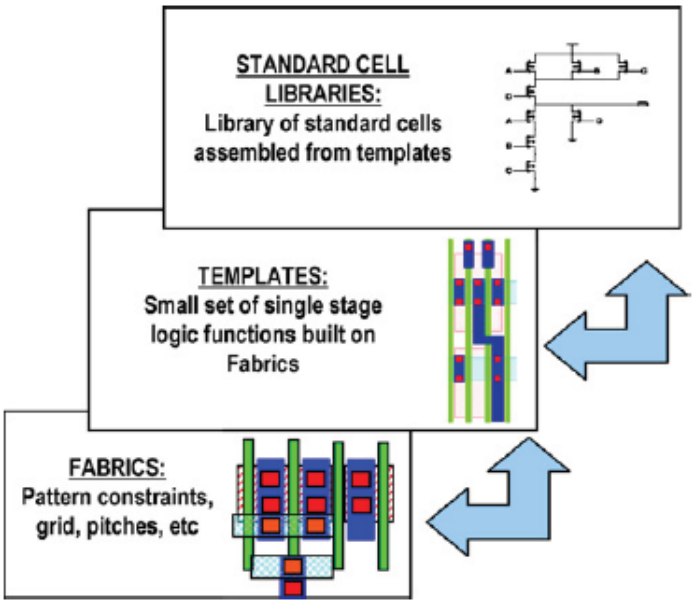


Figure 3.4: Flow for standard cell library creation from a regular design fabric [41]

The quest for highly regular and manufacturable fabrics and templates is still open. Lithography variability is the only source of variability considered so far and there is a need to consider the effect of other sources of variation. According to [40] there is a cost of regularity in terms of performance and area. There is a need to study the cost of regularity. Also we think that the design of fabrics is still a matter of art and there is a need to find a systematic and automated flow to generate the fabrics and the templates. A major challenge in this technique is handling layout systematic variation. Layout systematics are dominated by effects that have a large spatial range. This justifies the need to have a quantitative method to measure regularity and link this to manufactruability to be able to assess what level of regularity is sufficient. In the next chapter, we will be deriving a quantitative regularity metric for this purpose.

3.3.3 Variation-Aware Placement

Various work has been reported for variation-aware placement. Gupta et al. [43] proposed a timing optimization approach that exploits the opposite lithography-induced gate length variation experienced by dense and isolated pitches to compensate for each other. Detailed placement to improve leakage using through-pitch variation was proposed in [44]. The authors modified the placement of cells in small windows such that contexts that reduce leakage are created. Process variation due to lens aberration can be modeled for purposes of analysis and optimizations in the design phase. Kahng et al. [45] presented a timing analysis flow, that utilizes Zernike coefficients that quantify aberration along with layout information, to perform a more accurate analysis and reduce timing guardband. Then, they proposed 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. A detailed placement approach to reduce CD variation is proposed in [46]. Placement affects the pitches of devices in a layout, which determine CD variation arising because of proximity effects. Three algorithms, namely, cell flipping algorithm, single row optimization approach and multiple row optimization approach, are proposed to tune any existing cell placement to be lithography friendly. These algorithms are based on dynamic programming and graph theoretic approaches, and can provide different tradeoff between critical dimension (CD) variation reduction and wirelength increase. All these methods required a pre-characterization of libraries in various contexts. And this pre-characterization requires accurate models that are based on a stable process. By the time libraries are characterized, it is difficult to provide an accurate basis for models early in a technology node. There is a need to include more process effects to these methodologies, and a placement optimization approach will require to include various process effects concurrently.

3.3.4 Variation-Aware Routing

The work in the variation-aware routing seems to be focused on functional yield issues and solving the so-called process "hot-spots" problems, with some work being done targeting electrical variability and parametric yield.

Hotspot-free routers

There are numerous work that targeted a *Hotspot-free* routers. The use of two-dimensional pattern matching engine integrated into the router in order to enforce extra design rules to eliminate hotspots was proposed in [47]. The integrated approach allows rapid identification of hotspot patterns and allows for rapid fixing and verification of these hotspots by a tool that understands design intent and constraints. Mitra et al. [48] propose a lithography-aware routing technique that guides an off-the-shelf router to minimize edge-placement-error (EPE). First, EPE for the layout is estimated using lithography simulation. In each routing grid cell, the cumulative EPE density is calculated, and grid cells are processed in decreasing order of their cumulative EPE density. Two routing modifications are proposed in the paper: (1) spreading of routing segments in the neighborhood of a large EPE routing segment, and (2) addition of blockages followed by **ripup-and-reroute**. Fast aerial image simulation is also developed to monitor the impact of routing modifications on EPE. The authors found insertion of blockages followed by ripup-and-reroute to be effective at EPE reduction and report the associated EPE reduction to be up to 40%.

In [49] the authors propose an Efficient Lithography-Aware Detailed Router (ELIAD) that targeting post-OPC image in a correct-by-construction fashion. The router is configured to optimize post-OPC silicon image as part of the calculated cost function. In their formulation, they adopt a proposed lithographic-metric in ELIAD by applying a Lagrangian relaxation technique. Experimental results on 65-nm industrial circuits show that ELIAD outperforms a ripup-and-rerouting approach such as Resolution-enhancement-technique-Aware Detailed Routing [48] with 8X more EPE hot spot reduction and 12X speedup.

Electrical variability-aware routers

Cho et al. [50] propose global routing that accounts for topography variations. The authors observe that interconnect height increases, and consequently its resistance decreases, as the wire density decreases. Also, the coupling and total capacitance decrease with wire density. Thus, for timing-critical nets, it is beneficial to have low wire density in their neighborhood. The proposed router essentially reduce wire density in the vicinity of timing-critical nets

to improve their speed, and reduces wire density of high-density grid cells to reduce overall CMP variation. With the proposed approach, the authors claim a reduction of 8% in the minimum clock cycle time with negligible wirelength increase.

With the introduction of double patterning, there is an increase in the research work on taking the effects of variability of litho-effects. It is reported in [51] that the overlay error can cause up to 23% variation on the coupling capacitance and 17% variation in the RC delay of Metal1 layer. Therefore, routers should consider the effects of such errors on both functional and parametric accuracies.

3.3.5 Variation-Aware Post-Layout Analysis

In this subsection, state-of-the-art works that try to correct some of the systematic process variation problems post layout will be reviewed.
Balasinski et al. [52] propose a methodology of manufacturability qualification for ultra-deep submicron circuits, based on optical simulation of the layout, integrated with device simulation. They defined maximum and minimum accepted printed contours that ensure that transistors drive and leakage currents, I_{ON} and I_{OFF}, are within specified limits. The maximum CD tolerance contour would define the minimum drive current I_{ON} and the minimum CD tolerance contour would define the maximum leakage current I_{OFF}. They considered CD variation caused from proximity effects and from masks misalignment, and when failing to meet tolerance they suggested to choose among the following options:

- reduce spec limit for drive current (i.e., modify product parameters),
- change transistor model (change mfg process),
- reduce OPC hammerhead (risk: higher leakage), or
- change (tighten) outer tolerance contour (restrict exposure conditions).

The new model parameters should be verified and adjusted, until satisfactory solution is obtained We find value in their use of electrical parameters (I_{ON} and I_{OFF}) extracted from simulated contours to check if the design is meeting expected tolerance or not, on contrast

to using fraction of CD, but only accounting for limited sources of variation (proximity effects at nominal process and masks misalignment) limits the value of this proposal. Also we find the proposed suggestions for fixing when detecting problems are all limited to process change or change in product specifications. Both are not a real "design" actions.

Pack et al. [53] propose to incorporate advanced models of lithographic printing effects into the design flow to improve performance verification accuracy. They extracted the effective channel length of the transistors using two techniques: the faster gate averaging technique and a gate slicing technique. The extracted dimensions were annotated back to the circuit netlist and then used in SPICE-like simulator to study the effect of defocus variation on the timing of the circuits. While doing so, they compared different manufacturing technologies to see which technology will result in better performance. While this approach maybe useful in analyzing small circuits, it is not intractable to larger circuits - where digital designers would not use SPICE-like circuit simulators to analyze their circuits. Also this work did not propose a way to improve the performance or mitigate the variation effects.

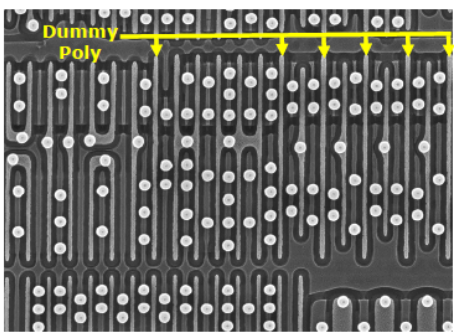
Orshansky et al. [54] studied intra-die gate length variability in a 180nm process and reported systematic variation to be more significant than random variation. Further, the authors observed spatially-correlated variation to exceed context-dependent variation that arise due to proximity effects. They also developed a theoretical framework allowing explicit analysis of circuit speed degradation due to L_{gate} intrachip variability. The observed extent of gate length variation induced a 25% variation on clock cycle time and the need for a systematic variation-aware timing analysis methodology was highlighted. The authors used a simple relationship between the gate length and the cell delay, and proposed a location-dependent timing analysis flow that accounts for spatial gate length variation.

Yang et al. [55] address post-lithography based analysis and optimization, proposing a timing analysis flow based on residual OPC errors (equivalent to lithography simulation output) for timing-critical cells and their layout neighborhoods. From the estimated gate lengths of all the devices in timing critical cells, the SPICE netlists of the critical cells are modified with the estimated device gate lengths, and standard-cell characterization is run. The critical cells are then mapped to the appropriate cell master in the library, and timing analysis is run. The authors report considerable change in slacks of several critical paths. It

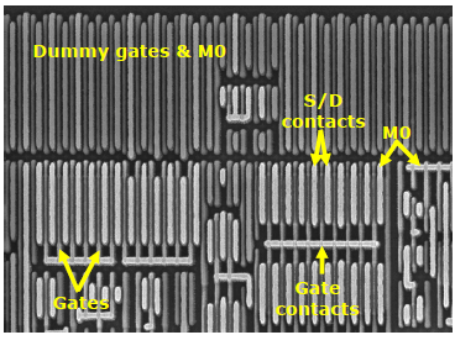
has been shown that ignoring post-OPC variation leads to under-prediction of the average slack by 24% and the worst slack by 36% in a modern microprocessor block. Path ranking in terms of their criticality is also significantly impacted. As a result, both the parametric and functional yields are potentially affected. Though this work showed the importance of considering post-OPC variation, only setup-time analysis was performed and interconnect variation ignored. Also, only nominal process condition analysis was performed. Several non-trivial details related to handling non-rectangular gates in SPICE simulations and cell-level hierarchy reconstruction are missing.

Full chip litho-simulation followed by a device level simulation may be accurate, however, a re-simulation approach is not compatible with the currently used timing flows. Standard flows rely on pre-characterized cell timing information for fixed cell footprints, and currently there is no easy way of using a post-OPC layout within a cell-based STA to perform delay estimation more accurately. In order to update pre-characterized cell timing information for a given cell that was printed in a specific manner new data models and tagging strategies need to be used. Re-extraction of layout parasitics and re-simulation of each cell based on the actual silicon profile generated by the litho simulation is too expensive in terms of STA runtime. Instead, parameterized cell timing models dependent on 2-D layout features may be used. The models will relate deviations of key geometries to changes in cell timing. Such models can be constructed using the technique of response surface method.

Gupta et al. [56] observe that lithography simulation permits post-OPC (optical proximity correction) estimation of on-silicon feature sizes at different process conditions. They propose a cell-level analysis flow that allows standard analyses tools to be used. After lithography simulation, cell instances of the same cell differ and cannot be mapped to the same cell in the library for lithography simulation-based analyses. Variants of each cell are added in the library; the variants are similar in function and drive strength of the cell but have different gate-lengths assigned to the devices. After rectilinearization and determination of gate-length of all devices in a cell instance, the variant that matches in the electrical behavior of the cell instance is selected and mapped to. The output is generated in the form of a modified Verilog file and can be used by standard analyses tools. Interconnects are simplified to polygons and their resistance is computed using analytical formulas.



(a) A 65nm AMD Athlon processor



(b) A 45nm INTEL XEON processor

Figure 3.5: Production Processors showing dummy poly and dummy gate [58]

For capacitance computation, pairs of interconnects are simultaneously simplified and the change in their coupling capacitance estimated using a pre-created lookup table. The same parasitic extraction approach is used to compute parasitics for corresponding drawn shapes and the change in parasitics is computed. The parasitic database is then updated with the change.

Cao et al. [57] also propose a full-chip timing and power analysis approach based on lithography simulation. In their approach, dummy features are inserted within a cell layout along the boundaries to shield from proximity effects. If on insertion of dummy features the proximity effects can be assumed to be negligible, all cell instances in a design experience identical lithography variation. Thus, no additional cell variants are needed in the library, and the cells can be characterized to account for the impact of lithography variation. The authors report 8%-25% reduction in timing guardband and 55% reduction in power guardband with respect to traditional corner-based analysis. For an industrial low power design, over 300ps reduction on the path delay variation was obtained.

James, D. [58] has reverse engineered some 65nm and 45nm parts manufactured in recent years, and examined extra structural features apparently added for DFM purposes. These are essentially extra lines of polysilicon as shown in figure 3.5(a), and in some cases STI blocks, added to equalize stress and improve lithographic uniformity as shown in figure 3.5(b).

Gupta et al. in [24] address the timing analysis implications of systematic variation in across-chip CD that arises due to imperfect defocus. Error in device CD (i.e., gate length) can be modeled once the defocus and pitch of the device are known. Gate delays depend on the CDs of the constituent devices, and the impact of across-chip CD variation on timing can be modeled. The timing analysis methodology proposed in [24] constructs variants of all cells in the library corresponding to different neighborhood contexts. In a placement for a cell C_i, its environment is described by a set of four spacings: nps_i^{LT} (distance of the device on the "left-top" to the nearest poly feature on the left in the neighboring cell), nps_i^{RB} (distance of the device on the "right- bottom" to the nearest poly feature on the right), nps_i^{LB} and nps_i^{RT}. These four space parameters enable us to determine the printed CD for the border poly features in the cell in the placement context using the through-pitch CD simulation results. They used three different values for each of these parameters. This gives rise to 81 different versions of the same cells. The appropriate variant is then selected from the library on the basis of the layout context of the cell to run timing analysis. The authors report a reduction of up to 40% in the timing guardband with respect to traditional corner-based analysis in static timing analysis.

Two approaches to creating variability-aware timing models for standard cells have been introduced in [23]. For both approaches it is assumed that all transistors within a cell experience identical process effects. The first approach utilizes geometrical biasing of transistor L and W in standard cells to create delay sensitivity tables. The second approach combines rigorous process simulation with contour based timing characterization to develop compact parameter delay models. Both approaches can complement existing standard cell characterization techniques. The authors claim that the delay responses for standard cells exhibit Bossung-like behavior, and are visualized in an electrical process window. This contour based timing characterization technique is applied to standard cells to investigate focus exposure variation, corner rounding, and layout proximity effects. Variability-aware timing models for standard cells in the form of delay variability tables or compact parameter timing models are shown to enable static timing analysis tools to perform variability-aware delay analysis on critical paths with little expense in runtime.

Variation in interconnects (a.k.a back end of line (BEOL) is also systematically analyzed in number of work. Sylvester et al. [59] observe that up to 60% of BEOL guardband can

be eliminated by use of the realistic BEOL variation model. Zhou et al. [60] propose a methodology that performs lithography simulation on the interconnects prior to parasitic extraction. The focus of their work is on construction of extraction rule decks using a 3D field solver for shapes outputted by lithography simulation. The impact of topography on interconnect parasitics is more extensively studied by He et al. [61]. Using the topography model developed for Copper CMP, the authors estimated the change in parasitics with 3D field solver simulations. For a 1mm interconnect in 65nm technology, the authors report the resistance to increase by 30% when CMP-induced copper dishing is accounted for. Capacitance impact was relatively small and typically under $\pm 3\%$ for coupling capacitance, and $\pm 0.3\%$ for total capacitance. However, with the insertion of fill, coupling capacitance increases by 30% to 140%, and total capacitance is impacted by -1.35% to 1.88%. The authors also propose a dynamic programming-based simultaneous wire sizing and buffer insertion algorithm that accounts for changes in parasitics due to fill insertion and post-CMP topography. With respect to traditional buffer insertion and wire sizing that is oblivious of CMP and fill effects, the proposed approach improves delay by 1.6%.

Post-Layout analysis is a very important area, but these techniques tend to be computationally intensive. There is a need to do this selectively at areas that are of interest to the designer. Since multiple sources of variation have correlated effect on performance, there is also a need to integrate all the sources of variation in a single analysis flow so that all effects are considered during the analysis. Also device level extracted parameters (for example variation in transistors V_{th} or L_{eff}) have little indication on the performance penalty of the variation - giving the designer little in-sight of where to focus his/her efforts to mitigate (or ignore) these variation. Integrating model-based layout legalization into a design flow still requires substantial reengineering of the entire IP generation, synthesis, placement, as well as routing flow and any attempt to do so must be careful not to increase the design cost, complexity and time. The most obvious un-answered question is how to react when the analysis reports a problem. The optimal design fix to legalize the layout is not always trivial. It requires a detailed understanding of the manufacturing process to find a modification that will find and legalize the layout as well as insight into the design purpose to ensure that design intent is not lost. Being so-late in the design cycle, any attempt to change would require several verification and correction iterations even during

the Place-and-Route step.

3.3.6 Post-Tapeout Variation-Mitigation Techniques

These techniques are not targeting the designers but rather a design-aware process optimization. But we will briefly outline the state-of-the-art work to highlight that one of the still unanswered questions in this approach is the lack of a vehicle to communicate the design-intent to the manufacturing facilities.

Selective gate-length biasing for leakage control was proposed in [62]. It is well known that leakage power decreases exponentially, and delay increases linearly, with increasing gate length. Thus, it is possible to increase gate length only marginally to take advantage of the exponential leakage reduction, while impairing performance only linearly. From a design flow standpoint, the use of only slight increases in gate length preserves pin- and layout-compatibility; therefore, the technique proposed in [62] can be applied as a post-tapeout enhancement step. They applied gate length biasing only to those devices that do not appear in critical paths, thus assuring zero or negligible degradation in chip performance. Selective gate length biasing at the circuit level reduces circuit leakage by up to 30% with no delay penalty. Leakage variability is reduced significantly by up to 41%

Several work proposed to modify the objective of OPC to minimize the electrical error, rather than edge placement error. An algorithm minimizing the difference in saturation currents of the contour and target shape was proposed in [41]. And Teh et al. [63] defined a transistor-performance-error (TPE) metric rather than the conventional edge-placement-error as the cost function optimization of OPC. Within the same concept of design-aware process optimization, the authors in [64] proposed to optimize the exposure dose map to optimizer timing and leakage.

3.4 Conclusion

In this chapter, the various techniques to combat systematic process variation in the design space have been reviewed. And as the analysis and mitigation techniques are proposed in

different levels of the design flow, there are still open questions on how to integrate these different alternatives. There is a need to provide the design community with an easy way to analyze and fix process variation issues and assess the cost of this fix. Post-Layout analysis techniques are the most efficient to handle systematic variation problems, but these techniques tend to be computationally intensive. Regular standard cells and regular fabrics designs are shown to improve the resilience of circuits to process variation, but there is still a problem in quantitatively measure regularity to compare different design techniques without the need for rigorous simulation steps. In the following chapters, we will develop fast and accurate models to relate process variation to electrical variation. And implement a framework that can abstract the complexity of the process and communicate the design constraints to lower levels of abstraction and we will propose techniques targeting to identify and mitigate both electrical variation (parametric yield) for logic blocks and critical failure hotspots (functional yield) in routing interconnects. The proposed techniques will aim to be faster and efficient compared to the previously reported ones.

Chapter 4

Regularity Metric to Model Electrical Variations in Logic Blocks

4.1 Introduction

As we have seen in the previous chapter, layout-induced process variations are the major contributors to the systematic die-to-die and with-in-die variations. Assessing the impact of systematic variation requires accurate, yet abstract models. The published work for modeling lithography variations (e.g., [65]) and stress variations (e.g., [66]) are based on capturing the physical changes in the devices, the device dimensions L and W in case of lithography and carrier mobility μ in case of stress. This requires designers to perform layout physical extraction (LPE) and model based extraction (MBE) to annotate the SPICE netlist with layout and BSIM model instance parameters. Then they would have to run SPICE simulation on the extracted netlist to measure the impact of the variations. In our research we will include device modeling to the physical variation modeling so that our models will relate back to designers parameters that actually represent their circuit performance without the need for subsequent circuit simulation. We will mainly target the ON and OFF currents (I_{on} and I_{off}) as they measure for timing performance and leakage power.

In order to provide optimized solutions for systematic intra-die variation, the design community needs a framework that allows smooth transfer of process variation information across multiple levels of abstraction of the design. The framework should allow accurate modeling of the physical variation effects, and feedback solutions to the various stages of the design. For the design iterations to reach a closure, solutions will be limited to the last stages of the design that are also the least perturbing stages. The design information will be used to smartly limit the computationally intense analysis to the areas that have the most impact on the yield.

In the remaining of this chapter, a novel method to model electrical variations due to systematic lithographic variations will be presented, moreover, a framework to link design information to the physical design (and vice versa) will be shown. The process variation caused by lithography and stress effects in a standard 45nm technology will be studied, and by calculating the effects of lithographic and stress variability on the electrical performance of the circuits using the developed model and the implemented framework we can gauge the importance of the accurate analysis and model-driven corrections. Based on the findings above a geometrical-based layout regularity metric is derived. This metric can be used as a fast indicator of designs more susceptible to process variations and hence electrical variations. The validity of using the regularity metric to flag circuits that have high variability using the developed electrical variations model is shown.

4.2 Process Variability Modeling

In this step we will build and use simplified but accurate models to model the effects of the major systematic proximity based process variations effects. We will focus on lithography variations and variations because of stress effects. These sources of variation were described in previous chapter (sections 2.3.1 and 2.3.4 respectively). The goal is to convert the variations in process parameters into variations in the electric parameters of the devices. In chapter 2, it was shown how fixed corners device models are over pessimistic. Although statistical device models provide a more realistic evaluation of the designs but still they treat variation parameters as random variables and do not benefit from the ability to

accurately account for systematic parameters variation. For this reason, our work will aim to derive accurate models that can directly model the systematic variations.

Unlike already existing models, the proposed models aim to spare the designer from running circuit simulator step on extracted parameters. For example the model should be able to predict the variations in current (on-current and leakage current) of a certain device under process variations without the need for computational expensive circuit simulation. The accuracy of the proposed models should be compared to the accuracy of the already existing methods where physical extracted parameters (e.g. effective channel length, effective mobility, ... etc) are fed back to the circuit simulator.

4.2.1 Lithographical Non-Rectangular Gate Modeling

Lithography simulation enables estimation of CD variations at different process points. A substantial fraction of variations is systematic and can be modeled accurately after layout. So even though random variations cause differences between on-silicon shapes and those predicted by lithography simulation, these differences are relatively small [56]. Yet litho-simulation is not enough, the challenge is to transform shapes generated by lithography simulation to a form that preserves the electrical properties.

Various recent work address the problem of non-rectangular gate (NRG) transistor modeling. The work in [67] proposed a method where each device is represented by a series of MOSFETs connected in parallel, each of them with the channel length determined from the litho-simulation. This approach has two limitations: the first is that the extracted transistors count would increase significantly as each transistor would be represented with multiple transistors and this would increase the complexity of the extracted circuit dramatically. Secondly, the standard BSIM transistor model would not work for thin slices because there is no compact model for transistor slices. This would negatively impact the accuracy of this approach. To overcome these limitations [68], proposed approximating NRG transistor with an effective length that is a weighted average of all the corresponding slices. The weights of the slices are proportional to simulated slice current. The slices current are extracted from a look-up table that is built from simulating very wide transistors with various channel lengths. The use of wide transistors in building the look-up table is

intended to eliminate the effects of short-channels. To calculate the current through a slice just multiply the current through the wide transistor by the ratio of the width of the slice or rectangular transistor to the wide transistor width. The drawback of this method is that it produces two equivalent devices: one when the transistor is ON for timing simulations and the second when the transistor is OFF for leakage simulations.

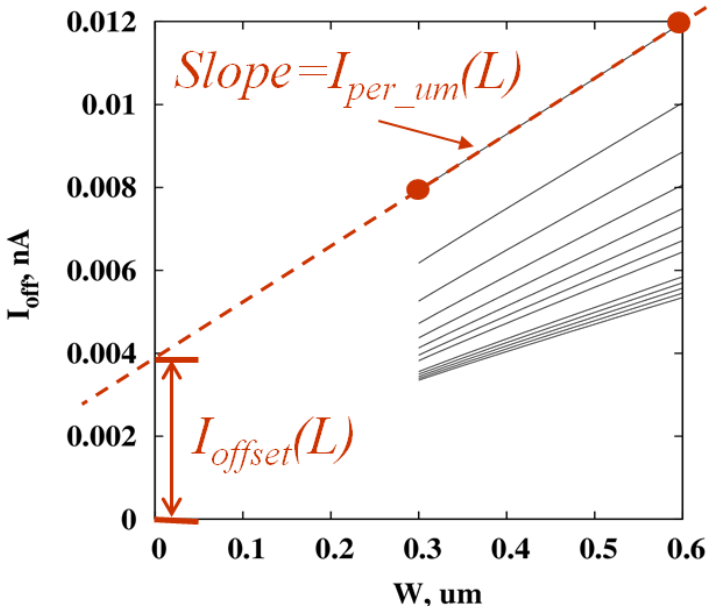


Figure 4.1: Transistor current for different gates width (W) and length (L)

We used the same method as in [65], where the current of rectangular transistor is approximated as:

$$I(W, L) = I_{per\_um}(L)xW + I_{offset}(L) \tag{4.1}$$

Where I_{per_um} and I_{offset} are obtained from transistor characterization shown in Figure 4.1. The coefficient $I_{per\_um}(L)$ denotes the slope of the line in the I-W plot and $I_{offset}(L)$ denotes the y-intercept of the same line. In this way, we can express $I(L, W)$ for all rectangular devices by computing $I_{per\_um}(L)$ and $I_{offset}(L)$ over a range of L values. So this would mean that $I_{per\_um}(L)$ represents the current produced per unit width from a very wide device, and $I_{offset}(L)$ represents the current offset due to effects which are not proportional to W. For NRG, the gate slicing method is used to compute the current of

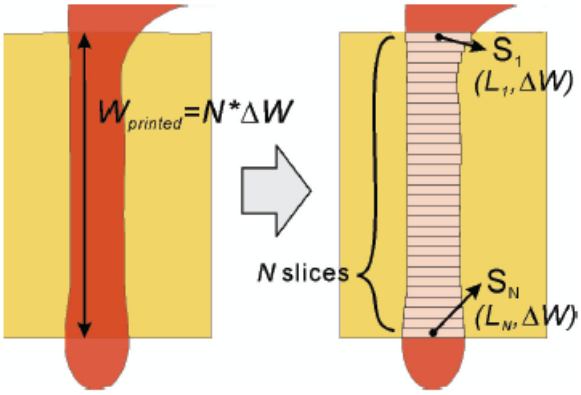


Figure 4.2: The NRG device contour is broken into parallel slices

the NRG device (Figure 4.2) as the sum of all slices, where the current for each slice is calculated as:

$$I_{slice}(L_i) = I_{per\_um}(L_i)xW \tag{4.2}$$

Considering the non-linear effect of narrow width ; we modified the gate slicing method to include an extra term which compensates for such non-linear effects for the edge slices. Since there is no need to compute the equivalent gate dimension because our method of directly calculating current from non-rectangular gates provides a quick and easy electrical performance analysis without SPICE simulations. The current is calculated from Eq.(4.3).

$$I = \sum_{i=1}^{n} I_{per\_um}(L_i) \times W + 1/2\left[I_{offset}(L_1) + I_{offset}(L_n)\right] \tag{4.3}$$

When compared to TCAD, this method had an average current error of only 1.6% [65] and had a correlation of 0.98 with measured transistors current [69]. This calculation was performed on each process condition and nominal, min and maximum ON and OFF currents of each transistor were calculated. Our method didn't need to compute the equivalent gate dimension because it directly calculates current from non-rectangular gates, providing a quick and easy electrical performance analysis ***without SPICE simulations***.

4.2.2 Stress Modeling

To calculate the effect of stress on the performance of the transistors, a commercial simulator (Mentor Graphics, Calibre) that predicts stress everywhere in the layout caused by a variety of sources, including stressed liners, epi-SiGe structures confined in the source/drain regions, tensile VIAs and STI, was used. These sources are located inside a floating window surrounding each gate that would extend up to 4000nm from each side of the transistor. The calculated stress is then used with a pre-calibrated model to calculate the change in the drive current caused by both mobility and Vth changes due to stress. The model is based on fitting extracted parameters from the layout to a piece-wise approximation of stress equation. The channel length, S/D diffusion length and STI width are extracted from the layout and stress and mobility are calculated from a pre-calibrated model [66].

4.3 CAD Framework

A methodology to link design information down to the physical design (and vice versa) is needed. Hence one will be able to map the variability induced by process variations to the parametric yield of the design. We will focus on risks in meeting timing and power specifications of the circuit, and identify areas in the design that will jeopardize meeting these specifications. The models generated in the previous step will be used to account for the process variations effects. In order to perform the analysis, it is required to integrate the physical design and its electrical connectivity in the same database with the design intent information. In addition, there is a need to model the effect of process variations on the circuit performance. Without such integrated framework, it is not easy to bring together the information gained from layout analysis, layout-aware circuit analysis, resolution enhancement and optical proximity correction tools, parasitic extraction, timing estimates, and stress analysis to suggest the DFM solution which is optimized within the existing constraints on design time and available data. The integrated framework as described in [70] describes a platform to integrate all the gained information. The variability analysis models and the fixing methodology will be built on top of this framework.

Having such an integrated framework will enable to transverse different levels of the

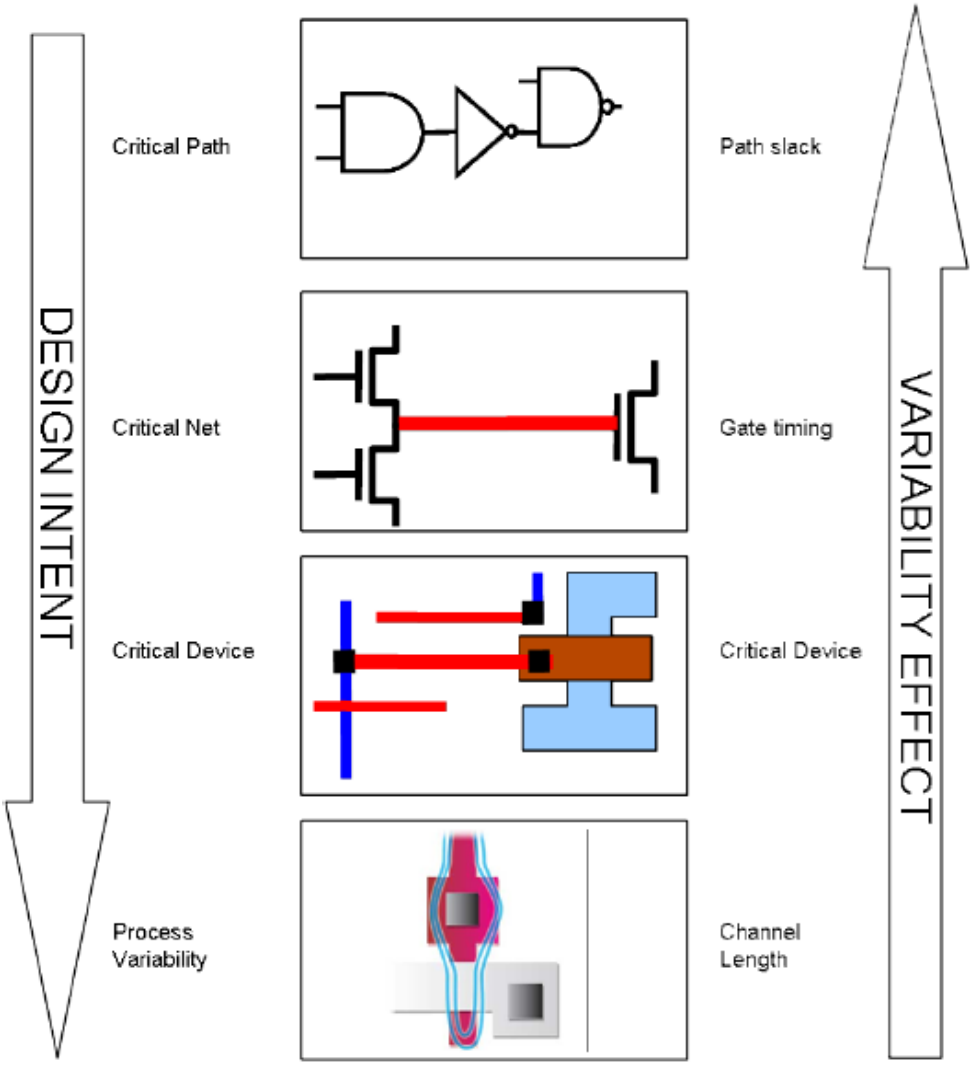


Figure 4.3: Design and process variations information flow

```
1: procedure FindSensitive(Design,CriticalPath)
2:     for each net N in Design do
3:         if N ∈ CriticalPath then
4:             Find devices T_List connected to N
5:             for each device T ∈ T_List do
6:                 Calculate Process Variation Induced Electrical Variation ΔI for T
7:                 if ΔI ≥ threshold then
8:                     add T to output_list
9:                 end if
10:            end for
11:        end if
12:    end for
13:    Return output_list
14: end procedure
```

Figure 4.4: Find Critical and Sensitive Devices Algorithm

design. As shown in Figure 4.3, design intention can be pushed down the levels of abstraction, so process variations analysis will only be done in areas of significance. With applying the proper modeling of process variations the effect on circuit performance can be propagated upward in the design levels.

We implemented the CAD framework that allows us to link the different design information from different levels. The framework accepts the variation-aware models described in the previous section. The integrated framework is used to find all the transistors in the layout on the critical path automatically and then calculate the process variation induced electrical parameter variations using the process variation models. Figure 4.4 shows the outline of the algorithm used. The algorithm of locating the "critical" transistors can be described as follows:

1. Netlist extraction is performed on the layout.

2. The critical nets are found from the STA timing report and they are mapped it to the extracted netlist.

3. The critical nets are then annotated back to the layout.

4. All the transistors whose drain/source are connected to the critical nets are found.

5. The standard cells that contain any of the "critical" transistors are labeled "'critical"'

Then the framework applied the process-variation models on the design to highlight devices that are sensitive to process variation.

With this ability to transverse the design levels and integrate design and process information one can direct the available analysis capabilities at the parts of the design where the results of such analysis are critically important for the performance of the chip. Only a subset of the devices on a chip are sensitive to process variations, also only a subset of the devices on a chip are critical for the performance of the chip. Only the intersection of these two sets, both sensitive to process variations and critical for the performance are worth the thorough analysis and will affect the parametric yield of the chip. Rigorous analysis and fixing efforts should focus on these devices. We would also like to note that for most sensitive devices, the exact characterization of their parameters is not particularly important. This suggests that approximate modeling techniques can be used to speed up and simplify the simulation tools when they identify sensitive devices. Once the set of sensitive and critical devices is identified, accurate modeling of those few instances becomes necessary on this set only. In this step we will be utilizing static timing analysis (STA) tools (we used both Mentor Graphics's Olympus or Synopsis PrimeTime) to identify critical paths and layout-vs-schematic tool (Calibre-LVS) for device extractions and electrical connectivity establishment.

4.4 Layout Regularity Metric

The target of this section is to look for a methodology that can differentiate between regular and irregular patterns in a quantitative way. This is more oriented towards micro-regularity, so the regularity mentioned here, means the regularity of certain pattern or cell or part of the layout and not the regularity of the whole design. We are going to

show latter in section 4.6 how irregularity is closely correlated to variations induced in the process (specifically lithography variation). By developing a metric to measure how regular a design is, one can estimate how the design will be resilient to process variation. The regularity metric has the following possible usage scenarios:

- It can be used in pre-layout phase to help the designer create correct by construct patterns from the beginning.
- It may be used post layout in design verification phase to give something similar to critical area analysis where it highlights the problematic patterns that are more susceptible to process variations. The best practices deduced from the model will be used to provide hints to the designer to help him solve these problematic areas.

4.4.1 State of the art Regularity Metric Techniques

The use of two-dimensional Fourier transform to compare between different layout styles was proposed in [71]. The comparison was based on analyzing the number of dominant frequency components in the Fourier transform of each layout style. The regular layout that uses a small number of layout patterns placed at a fixed pitch is expected to have a high degree of repetition and thus, a finite number of dominant frequency components. This method provides a visual comparison of regularity but it does not give enough information to be able to compare accurately two layouts of somehow similar regularity. It can be used to compare regular versus non-regular layouts but it is difficult to use it to compare similar layouts in terms of regularity. One limitation in this method, is that it does not allow to highlight the locations of geometries or patterns that are causing the irregularity, but only calculate the frequency components. Also, it can be noticed that Fourier transform is computationally intensive and expected to have long runtime.

More recent work related to regularity metric was done in [72] on 65nm technology node. Layout regularity was defined, for a certain layout layer, as the ability to represent this design layer by a small number of constructs. According to this definition, the maximum regularity is achieved when a single construct can be used to generate the whole layer. On the other hand, the minimum regularity occurs when different unique constructs are used.

The Fixed Origin Corner Square Inspection (FOCSI) proposed in [72] first exports the layout layer as an image and then detect all upper left pattern corners. Then, considers these corners as the origins of the square grids to be compared sample by sample against each other in order to calculate the number of different constructs for each sample grid. FOCSI method seems to be better than Fourier Transform in the sense that it can provide more quantitative measure and can compare the regularity of two layouts created using the same technique. However, converting the layout to image and comparing it pixel by pixel is definitely compute intensive and will require long runtime too.

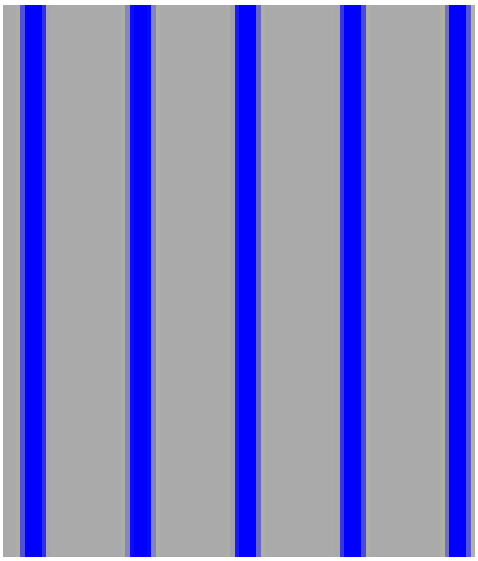


Figure 4.5: Most regular and least variable pattern

4.4.2 Geometrical based Layout Regularity Metric

For the purpose of simplicity and fast computational time, a geometrical based method is used to define regularity. Starting from the well-known fact that a pattern consisting of parallel lines with equal widths separated by equal spaces as that shown in Figure 4.5 is considered the most regular pattern and it has the least variability. This means that the regular pattern has: (a) single orientation, (b) regular density, and (c) regular pitch.

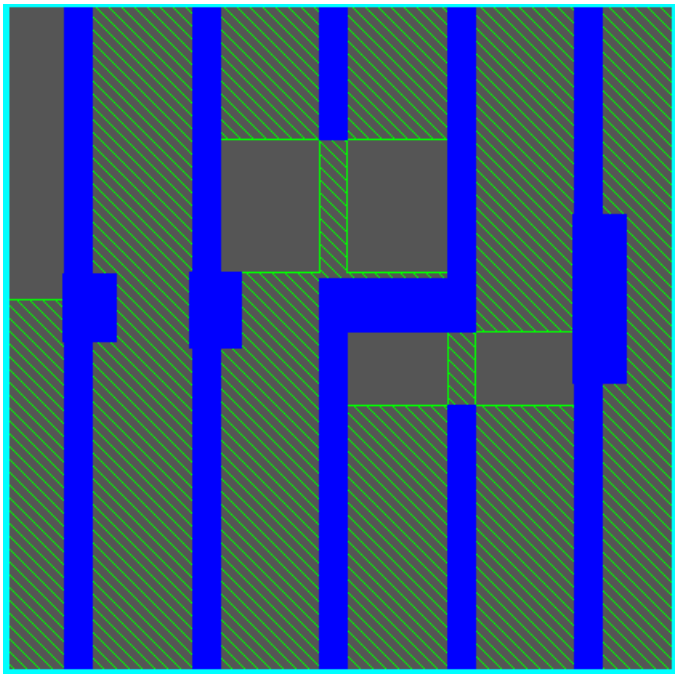


Figure 4.6: Example of derived layer for a certain pattern (hashed)

A metric was derived using a simple equation shown in Eq. 4.4. This equation contains geometrical properties in a way such that the metric has a maximum value when the pattern resembles the most regular pattern. The metric value decreases as the pattern has line ends, jogs, corners and shapes of different orientations and different densities.

$$RM \propto \frac{\sum \text{lengths of edges in favored orientation}}{\sum \text{lengths of edges in unfavored orientation}} \times \frac{\sum \text{perimeter (shapes of layer)}}{\sum \text{area (shapes of layer)}} \times \frac{\sum \text{perimeter (shapes of derived layer}}{\sum \text{area (shapes of derived layer)}} \tag{4.4}$$

Where RM is the regularity metric. derived layer is a layer created between the edges of projecting shapes within certain distance specified by the minimum spacing for each layer. Example of derived layer for a certain pattern is shown in Figure 4.6.

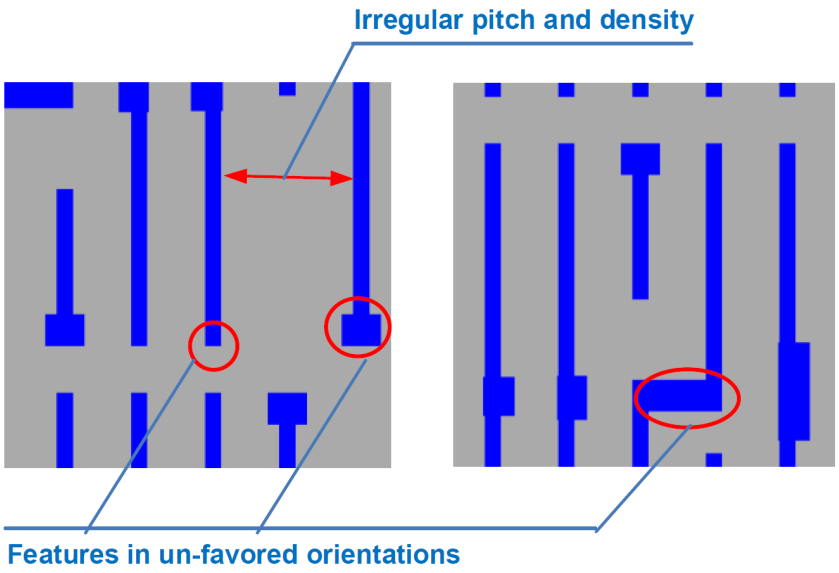


Figure 4.7: Sources of Irregularity

The regularity metric consists of three terms; the first term accounts for single orientation, the second term accounts for regular density and the third term accounts for regular pitch. Figure 4.7 shows two irregular patterns in the poly layer indicating the sources of irregularity.

4.5 Results

To validate the feasibility of the flow, device sensitivity to process variability is analyzed by studying both lithography variation and stress variation. A new simplified model to calculate the transistor current variation stemming from CD variations caused by lithography was developed. A CAD framework that allows us to traverse the different design levels, from logic gate level, down to device level is implemented. This framework allows the correlation of the circuit criticality with process variation. We show that only a small fraction of devices whose characteristics are significantly affected by process variability actually have correspondingly significant effect on the overall circuit performance.

In this work, setup time critical path is considered, and all the transistors in the cells on the worst critical path are tagged as "critical" transistors. The critical paths were identified by the static timing analysis on worst case timing. The outputs of the STA are the pins (nets) on the worst negative slack path. We then used our process variation models to study the effect of lithography and stress variations on these devices. Transistors with process-induced ON current-variations >= 10% relative to nominal current were considered as "sensitive" transistors.

The proposed flow was applied on three designs. The designs are implemented using an industrial 45nm technology. We analyzed the effects of process variations; both lithography effects and stress effects both lithography and stress models are calibrated to best match the silicon results. The lithography variations were modeled by simulating across dose and defocus variations. Change in dose of +/- 3% and defocus of 100nm were used. Calibre LFD tools were used to simulate the lithography step and generate full chip across process window contours. Stress effects were considered in a window of 4μm.

Design 1: S13207

Figure 4.8: The critical path of S13207

The first design is the S13207 design of the ISCAS'89 benchmark designs. This small block

has 23562 transistors. Out of this small block 50% of the transistors were process sensitive according to our criteria defined above. From these sensitive transistors only 77 (0.7%) are also critical in timing. 74 transistors of these critical and sensitive transistors are due to stress effects and only three of them are due to lithography effects. These 77 critical and sensitive transistors are in 16 instances of the 2484 cells in this small design.

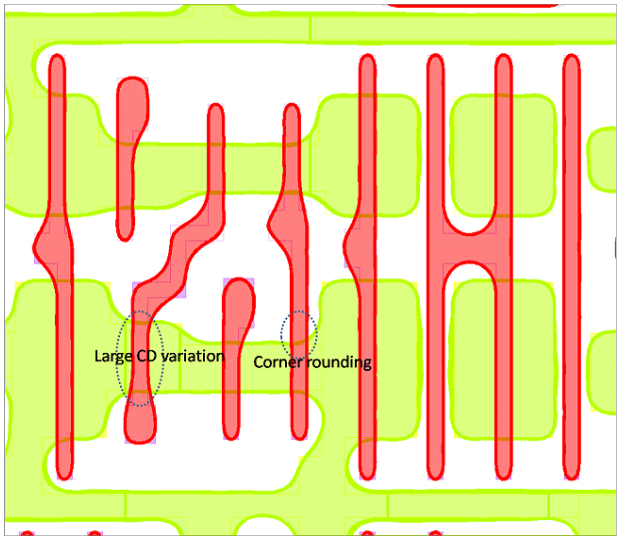


Figure 4.9: Lithography CD variations of one cell in S13207

Figure 4.8 shows the design's critical path mapped on the physical layout using the implemented framework to map STA results to the physical design. Figure 4.9 shows a snapshot of the simulation result, the thickness of the contour indicates the variability in critical dimension. It is obvious how the 2D irregularities cause more variation. The non-rectangular gate models was applied on the design and calculated the variability of the transistors on-current. Figure 4.10(a) plots the color map of current variations because of lithography variation.

When overlaying the critical path with the variability map results, the cells that has high variability and fall on the critical path are obtained as shown in Figure 4.10(b). In this and the subsequent experiments we conducted the simulation on the full design in order to validate our assumption that only a smaller subset of the design that requires thorough simulation. The proposed flow, is to only conduct the process simulation and

(a) Variability color map

(b) Critical and variable cells

Figure 4.10: Variability results for S13207

analysis around the designs critical path and to apply the fix only to these that are both highly variable and critical.

Design 2: b22

The second design is the b22 design of the ITS'99 benchmark circuits. This small block is composed of 62437 transistors, of which 47% are sensitive but only 93 are critical and sensitive.

Design 3: Digital Filter

This is a medium size (1mm x 1mm) commercial design, with more than 5.5 million transistors and around 300,000 cells. For this design we only considered lithography variations. Around 50,000 of the instances were considered sensitive, only 16 of them are both sensitive and critical. The set of critical and sensitive cells contains 8 library cells. We found two library cells (CK2D8 and INR2XD4) contributing to 9 instances in the critical and sensitive set. These two cells were also sensitive in most of their instances in the full design – indicating that these cells are sensitive to process variations regardless to their context. Other cells showed strong dependence on their context, this was deduced from their num-

Table 4.1: Distribution of critical and sensitive cells in the design

Cell	Critical	CandS	Sensitive	Layout
CKBD16	1	1	412	461
CKD2D2	1	1	228	228
CK2D8	6	5	307	307
DCCKBD16	1	1	66	71
NR2XD3	1	1	83	121
INR2XD4	3	3	355	496
NR2XD4	1	1	112	131
NR3D2	1	1	46	93
NR3D3	1	1	53	53

ber of sensitive instances compared to their total number. Table 4.1 lists different cells and their distribution in the critical and sensitive sets.

4.6 Analysis

In the previous section we have showed that a significant fraction of devices is affected by the layout context and should be considered sensitive. However, only a small fraction of these devices is critical for the circuit performance. This is especially true in large designs. Obviously, to make the design more robust we have to avoid devices which are both sensitive and critical. We would like to also note that for most sensitive devices, the exact characterization of their parameters is not particularly important. This suggests that approximate modeling techniques can be used to speed up and simplify the simulation tools when they identify sensitive devices. Once the set of sensitive and critical devices is identified, accurate modeling of those few instances becomes necessary.

4.6.1 Variability and Irregularity

From analyzing the results of variable devices, it was found that we can classify all the highly sensitive devices into three categories according to their geometries:

1. Non-uniform pitch as shown figures 4.11a and 4.11b, where the poly lines were not in the favored pitch by the dipole optical source, causing high variability in the gate dimension across process window.

2. Gates with incomplete coverage from neighboring poly lines, as shown in figure 4.11c, this was causing variation in the stress component and also variation in the gate dimension across process window.

3. Gates that are affected by neighboring irregular two-dimensional geometries.

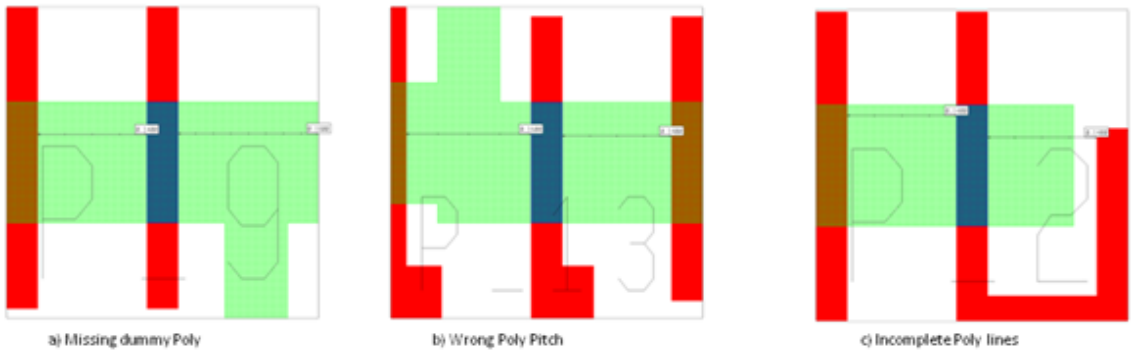


Figure 4.11: Layout dependent sensitive devices.

To verify the relation between the regularity and variability, the regularity metric as defined in section 4.4 was used to compare the cells that are found to be electrically variable to those found to be irregular and investigate the validity of using the irregularity as an indicator of highly variable cells. To define how the regularity part is obtained, the regularity metric value is calculated for the poly layer of all cells in the critical path, then cells are arranged in ten bins according to their regularity metric value and those cells in the least two bins are considered the most irregular patterns.

To be able to compare the results of the regularity versus that of the electrical variability, the same design benchmarks were used. Design b22 of the ITS99 benchmark circuit is composed of 7266 cells, 104 cells are on the critical path and of the 104, 49 were found electrically variable. S13207 design of the ISCAS89 benchmark designs has 2419 cells , 21 cells where on the critical path and out of the 21 cells , 16 were found to be electrically variable. Table 4.2 shows the number of cells on the critical path with electrically variable transistors according to the definition in previous section. It also shows the number of most irregular cells in the critical path. Then demonstrates the number of matches which are the variable cells that the regularity metric was able to detect, the number of misses which are the cells that were found to be variable but the regularity metric did not consider them highly irregular. The extras are the cells detected by the regularity metric as irregular while they were not found to be variable.

Table 4.2: Results of electrical variability and regularity results

Design	b22	s13207
Cells on critical path	104	21
Electrically variable cells	49	16
Irregular cells	33	14
Matches	31	12
Misses	18	4
Extras	2	2

To understand the misses and the extras in each design, we examined each one of these cells. The variability results showed that some of the cells placements were found variable and others were not, which means that these cells were affected by their neighborhood. A detailed study below clarifies this conclusion. The 49 electrically variable cells placements in design b22 were found to consist of 11 unique cells; the regularity metric detected 1 out of the 11 as irregular. In design 2, the 16 electrically variable cells placements were found to consist of 14 unique cells; the regularity metric detected 10 out of the 14 as irregular.

Misses are defined as cells that are regular according to the geometrical regularity metric yet they are shown to have high electrical variability. All 18 misses in design 1 were found

to be different placements of one cell shown in Figure 4.12 (a). According to the regularity metric, this cell has medium regularity. This cell had 58 placements in the critical path; 18 placements were found electrically variable while 40 were not. This can be attributed to the small size of the cell which makes it highly affected by neighboring cells. For design s13207, one of the four misses was the same cell as in Figure 4.12 (a). The other three misses in design 2 were different mirror images of the cell in 4.12 (a). 4.12 (b) shows an example of design s13207 misses.

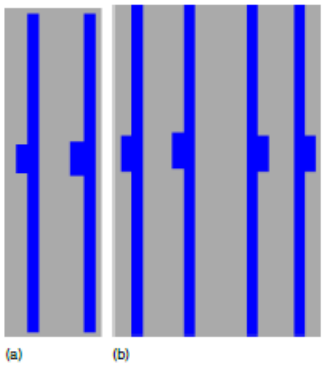


Figure 4.12: Misses: Regular cells that have high variability

Extras are defined as cells that have small electrical variability even though they are irregular. The extras in design b22 were two cells; one of them (4.13 (a)) had 24 placements in the critical path; 23 placements were found variable while only one was not. According to the regularity metric this was irregular cell. The other cell (4.13 (b)) did not have any other placements in the critical path but it was noticed that the gate region is far from the irregularities in the poly which may be the reason why it was not found electrically variable. The same cell in 4.12 (b) was one of the extras in design s13207 while the other one was a cell that looked a multiple of that in 4.13 (a).

4.6.2 Regularity trend in Advanced Technology Nodes

An important point to notice is that for 32nm and beyond, poly layers are becoming more regular. One can observe in Figure 4.14 [73] how the strive for uniform regular design has

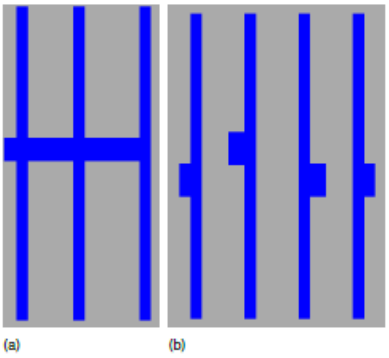


Figure 4.13: Extras: Irregular Cells that have low variability

progressed along the different nodes. The regular poly layers in 32nm and beyond indicate that the variability in the devices will be controlled and hence the complexity (and the problems) will be migrated to the higher layers in the stack.

4.7 Conclusion

In this chapter, a flow for analyzing the systematic process variation in nanometer CMOS technology was proposed. By calculating the effects of variability on the electrical performance of circuits we can gauge the importance of the accurate analysis and model-driven corrections. Lithography variation models were built, and an integrated framework was implemented to provide the design community with an easy way to analyze and fix process variations issues and assets the cost of this fix. The relationship between electrical variation and design regularity was established and we showed results demonstrating the flow on 45nm benchmark designs. We developed a metric for measuring the regularity of the design and demonstrated the ability of this metric to predict the sensitivity of the design to process variations. The metric can be used by designers to quantitatively assess the regularity of their designs and highlight areas of low regularity to fix. The metric can also be used to compare between different design styles. Designers using external IP designs, can choose to use the designs with the better regularity when comparing different IP vendors. In the next chapter we will use the relationship between regularity and variability that

was established in this chapter to identify critical lithography failures. In the following chapter we will introduce a method for fixing interconnect lithography failures based on the regularity metric we derived in this chapter.

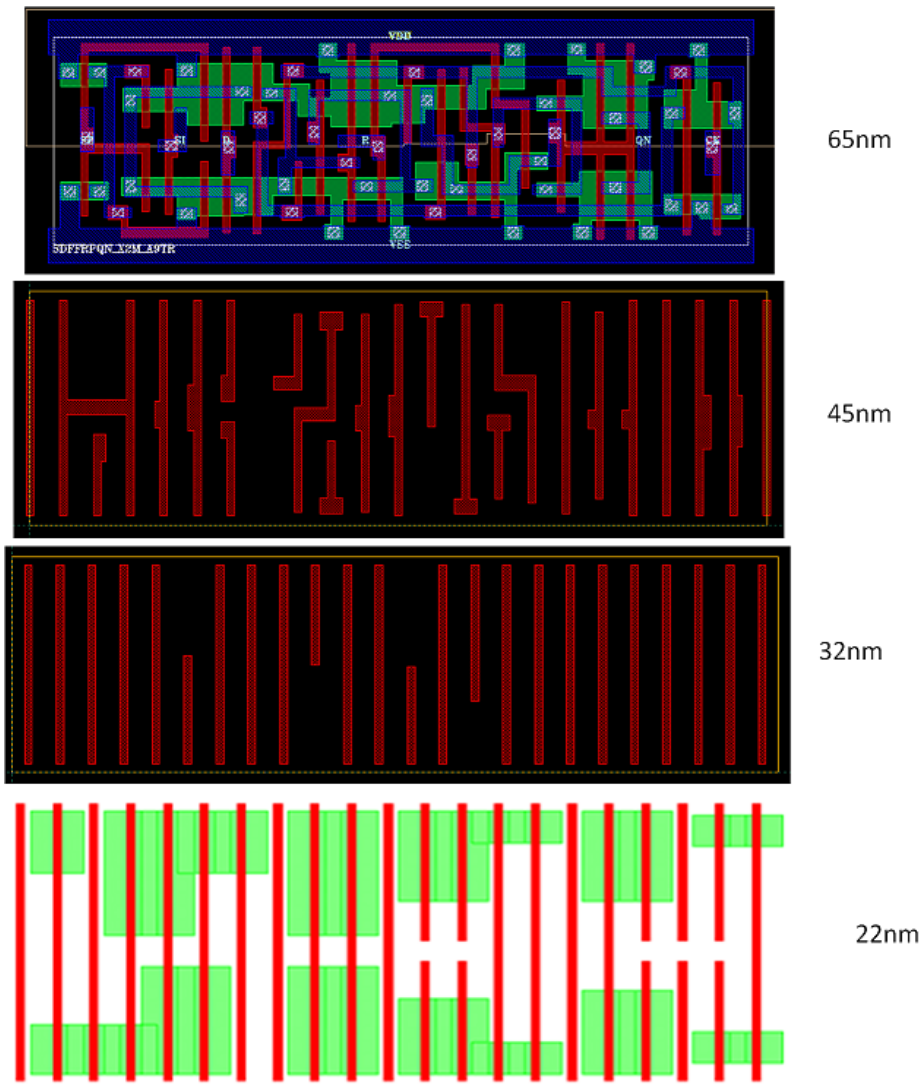


Figure 4.14: Standard cell design progress along technology nodes.

Chapter 5

Catastrophic Hotspot Detection using Machine Learning

5.1 Introduction

Lithography hotspots are layout patterns sensitive to lithographic process variations which degrade manufacturing yield. Hotspots need to be detected and fixed during the layout design and verification stages. Conventional lithography simulation [74] uses process models to generate the patterns shapes on silicon across the process window. Although it is accurate, the process simulation is a computationally expensive task. Pattern matching is a fast hotspot detection methodology [75] composed by a library of known problematic patterns and a search algorithm which identifies instances of the library elements in a layout of interest. Pattern matching is good at detecting pre-characterized hotspot patterns but has a limited ability to recognize previously un-characterized patterns. Recently, supervised machine learning techniques have been proposed to detect hotspots. [76–84]
In this chapter, a hybrid technique of machine learning and pattern matching is used for hotspot detection is proposed. The benefit of using a hybrid approach of pattern matching and support vector machines (SVM) is to maintain the detection accuracy of pattern matching techniques and maintain the predictability of the data learning systems to detect hotspots. The main contributions include:

- A hybrid pattern matching-SVM based flow that learns from known good and bad shapes, and builds a system to analyze layouts to identify hotspots.
- An encoding technique for patterns that is aware of the lithography problems and based on the regularity metric.
- A topological clustering technique to improve the accuracy and the adaptability of the system to change in manufacturing process.
- Data sampling of training data samples to overcome the problem of hotspot/non-hotspot imbalance.
- We apply our technique on the ICCAD 2012 benchmark data [85] and compare our results to other published techniques.

The rest of the chapter is organized as follows. First problem definition is stated in Sec. 5.2. Previous work in the literature is presented in Sec. 5.3. The proposed hotspot training and detection flow will be described in Sec. 5.4. Finally, experimental results and performance analysis will be shown in Sec. 5.5, followed by the conclusion in Sec. 5.6.

5.2 Problem Definition

The hotspot detection problem as defined in the ICCAD 2012 contest [85], is given two sets of layout clips that represent hotspots and non-hotspots sets; it is required to build a system that can identify hotspots in any layout. The training set is composed of layout clips similar to the one shown in Figure 5.1. For hotspot clips, there will be a center "'core"' square indicating where the hotspot appears. The remaining area of the clip, the "'frame"', indicates the amount of context area needed to introduce the hotspot inside the core area.

In order to validate the accuracy of the system, the results will be compared to process simulation results. A true hotspot that is detected by the system will be called a "'Hit"', while the hotspot that is not detected will be called a "'Miss"'. Falsely detected hotspot

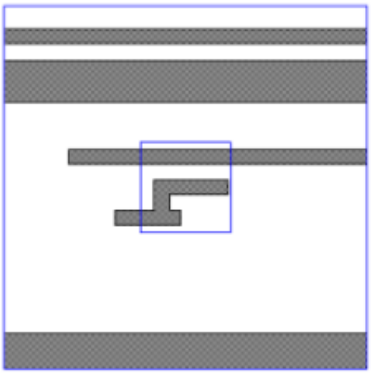


Figure 5.1: Example of hotspot pattern $frame$. Center square is the *core* area.

will be called an "'Extra"'. The target of hotspot detection systems is to have a high number of true hotspots (Hit) and low the number of false hotspots (Extra).
Expected targets for a hotspot detection system [85]:

- High Detection accuracy: $> 80\%$
- Low false alarm: < 100 false hits/mm^2.
- Fast run time: < 1 CPU-hr/mm^2

These targets are distributed along three axis including: Accuracy (Hit Count), False count (Extra) and Performance (runtime). High accuracy, means the system is capable of finding new hotspots that were not previously detected as part of the training phase. Having false counts (Extra), would result in an overhead where designers attempt to fix problems that do not exist. The objective is to have a system that bridges the gap between the two extreme methodologies of physical verification: Simulation based systems and Exact pattern matching based systems. Exact pattern matching is very fast < 0.1 CPU-HRS/mm^2 with low false counts but for a complete blind testcase the hit count can be zero as it fails to predict hotspots that were not seen before. Model based simulation is most accurate (golden reference which means hit count = 100% and false count zero), but with performance of the order of 100 CPU-HRS/mm^2. The objective of having a system

with accuracy greater than 80% and yet running at 1 CPU-hr/mm^2 (i.e. 100X faster than litho-simulation based system) is a reasonable target.

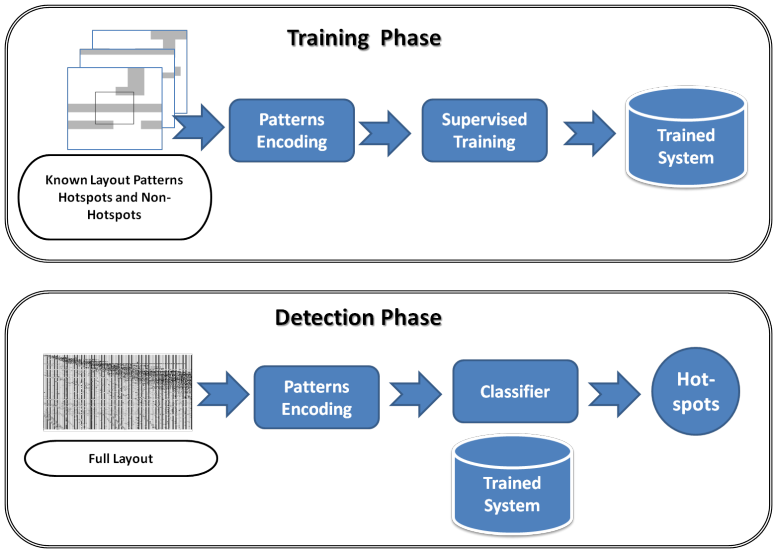


Figure 5.2: Basic hotspot detection flow.

5.3 State of the Art

Recently, several hotspot detection approaches have been proposed based on machine learning techniques to avoid CPU-intensive lithography simulations. Several techniques have been proposed to extract critical information of these hotspot patterns and be able to classify patterns with high accuracy and low false alarm. The basic hotspot detection flow in figure 5.2 is composed of two phases: "'The training phase"' where known hotspots and non-hotspots patterns are fed to the system and "'The detection phase"' where hotspots are detected in layouts. The training layout clips of known hotspots and non-hotspots are the inputs to the training phase. First the layout patterns in the clips are encoded in a way to extract the critical features in the patterns and represents them in a into a vector of real numbers. Then the training data set is used to train a supervised machine learning system. The trained system is the output from the training phase. In the detection phase,

the trained system generated in the previous training phase is used to find hotspots in new layouts. The input layout has to be scanned and its patterns should be encoded using the same encoding method used in the training phase. Then the encoded patterns will be classified to either a hotspot or non-hotspot according to the previously trained system. A training process is needed for each layer in each technology node. The training set will be obtained from experimental designs that are either simulated using accurate process simulators or from silicon data. Once the system is trained it can be used to detect hotspots in many layouts. In the following sections we will review the basic blocks of the state-of-art machine learning hotspot detection systems.

5.3.1 Patterns Encoding

An essential step for the hotspot detection in machine learning methods is to present layout patterns in such way that can describe the layout and make the task of classification easy. Several layout encoding methods have been proposed to extract critical features from the layout. The encoding transforms each layout pattern into a vector, an ordered list of real numbers. A good feature encoding scheme should properly represents the critical features that contribute significantly to the classification of the patterns.

Density of Pixels

The density-based pattern encoding is introduced in [76] and used in [77,78,81,86]. Figure 5.3 illustrates the basic concept. Given a layout pattern of a predefined grid, the method calculates the layout covering density of each grid. Then, the covering densities of grids of a layout pattern are encoded as an ordered feature vector. The ordered feature vector of the layout pattern is then mapped into a node in the multi-dimensional space.

Instead of the sliding window approach suggested in [76], authors of [77] and [78] suggested global grid across the entire layout for faster performance. With applying global grid, patterns are no longer guaranteed to be aligned on the same grids every time, so [77] proposed using morphed versions of the patterns by shifting each pattern half grid size in all dimensions to accommodate for all the maximum shifting in the grids.

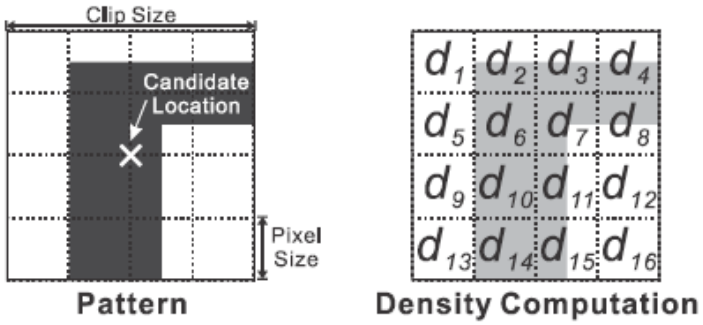


Figure 5.3: Density Based Pattern Encoding [78]

.

Topological Representation

The topological representation technique extracts a representative set of parameters from the design layout. These set of parameters are the most effective features to the presence of hotspots. The extracted parameters are mapped into metrics and data structures to represent the original pattern with significant run-time reduction. For effective representation, the mapped parameters should remain the same regardless to the orientation or the location of the patterns. Hann grids were used in [80] to generate fragments and Calibre's OPC engine was used to generate fragments in [79].

As shown in Figure 5.4, the context of each fragment F is defined as the neighboring fragments within a radius r. Only fragments inside this context are needed for a complete representation of the fragment. For each fragment F, the width of the polygon, the spacing to the facing polygon, the length of the fragment and the number of concave and convex corners are measured. Each pattern is then represented as an ordered featured vector composed of the critical feature measurements of all the fragments within the context radius.

An alternative approach to fragment based encoding, is critical features extraction [82]. This approach is based on extracting three features from each pattern. The Bounded Rectangles features are represented by 5 parameters: the length, width and orientation of the rectangle in addition to the coordinates of the upper-left corner. In addition to the bounded rectangles, T-shapes and L-shapes are also extracted and counted. A feature

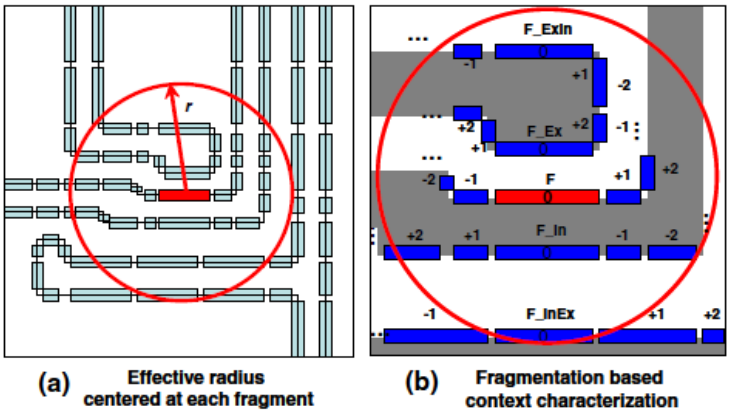


Figure 5.4: Fragment Based Context Pattern Encoding [79]

.

metric vector is derived for each bounded region in the clip in the form of width, length, orientation, x and y coordinates, T-shape count and L-shape count. The entire layout sample clip will be represented in the form of a sorted collection of these vectors.

Another technique is to combine topological critical features with lithography-process-related critical features of each pattern [83]. Four types of topological features are extracted: (1) horizontal and vertical distance between a pair of internally facing polygon edges, (2) horizontal and vertical distance between a pair of externally facing polygon edges, (3) diagonal distance of two convex corners, and (4) horizontal and vertical edge length of a polygon. Considering eight possible orientations, two sets of topological features are generated to preserve the vertical and horizontal relationships among extracted features. On the other hand, five types of non-topological features are extracted: (1) The number of corners (convex plus concave), (2) the number of touched points, (3) the minimum distance between a pair of internally facing polygon edges, (4) the minimum distance between a pair of externally facing polygon edges, and (5) the polygon density. Similar to the other techniques the ordered feature vector for each pattern is described by the values of the extracted features.

5.3.2 Supervised Training System

The problem of hotspot detection can be described as a binary classification. Supervised training Machine learning techniques are well suited for such a problem. Machine learning techniques construct a regression model (kernel) based on a set of training data. Many recent approaches utilize artificial neural network (ANN) and support vector machine (SVM) techniques to implement the hotspot detection kernel. Also a hybrid approach of using both SVM & ANN is presented in [84] to further improve the performance.

Artificial Neural Network (ANN)

ANN are computational models inspired from imitating human brain neuron networks and human learning activities. During training the ANN classifier, which is a neural network structure, calculates the outcome for the data sample vector and assigned weights and biases to the network are optimized to minimize the summed square error.
In [82], the authors feed the neural network with the extracted critical features for supervised training, which is an iterative coefficient updating to optimize all the neurons in the network. Then new unseen designs will be analyzed by the trained ANN kernel for hotspot detection tasks.

Support Vector Machine (SVM)

Support vector machine (SVM) is a statistical machine learning method used for classification and regression. SVM constructs a hyperplane or set of hyperplanes in a high dimensional space. The separation hyperplane, or decision boundary, is constructed such that the margin between the two classes is maximized.
The formulation of SVM [87] can be described as: Given a set of l training vectors $x_i \in R^n, i = 1, \ldots, l$, and a vector $y \in R^l$ such that $y_i \in \{1, -1\}$, SVM classifier (C-

SVC) requires the solution of the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i \\ & y_i(w^T\phi(x_i)+b) \geq 1-\xi_i, \\ & \xi_i \geq 0, i = 1,\ldots,l. \end{aligned} \tag{5.1}$$

where $\phi(x_i)$ is a mapping function for xi_i into higher-dimensional space and C >0 is the regularization parameter. The dual problem 5.2 is derived from the above formulation 5.1 and due to the possible high dimensionality of the vector variable w, SVM solves the dual problem.

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ & 0 \leq \alpha_i \leq C, \qquad i = 1,\ldots,l, \\ & y^T\alpha = 0, \end{aligned} \tag{5.2}$$

where e is the vector of all ones, Q is an l by l positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, and $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is the kernel function.

By solving the problem Eq.(5.2), the decision function is obtained to be:

$$f(x) = sign(\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b). \tag{5.3}$$

We store $y_i\alpha_i$, b, the support vectors, and other information such as kernel parameters in the model for prediction.
SVM classifiers are used for hotspot detection in several work [77, 78, 86]. Use of two levels SVM was proposed in [81] while [79, 80] extended this to multi-level SVM for false alarm minimization.

5.3.3 Complementing Machine Learning Systems using Pattern Matching Techniques

Pattern matching is widely used in the industry to detect yield limiting structures. A database of already known problematic patterns is constructed, then the pattern matching system scans through the layouts to find patterns that match the ones defined in the database. When comparing Pattern Matching to Machine learning [81], it is shown that pattern matching has a better accuracy in detecting already seen patterns in the training set, but has a poor predictability of unseen patterns. This inspired the authors of [81] to suggest a hybrid flow of pattern matching and machine learning to detect hotspots. A patterns database of "'known-bad"' shapes is first matched against the design. The matched locations are already known to be hotspots and so there is no need to consider them using the machine learning classifiers. The classifiers are then applied on the remaining, un-match, areas of the design. Finally, the results of the classifiers are combined with those from the pattern matcher.
An alternative hybrid flow is suggested [77], in which using pattern matching to detect all the outlier misses and false detections in each of the regions (based on the training set), which will be added or removed from the set of hotspots later on. Doing so allows: Reduce the number of patterns that need to be pattern matched since only the outliers of the machine learning system need to be considered and more importantly it allows addition of trained predictability to new configurations that were not in the training set but that can be interpolated from the system.

5.4 Proposed Flow

5.4.1 Overview

The hotspot detection flow is composed of two phases: "The training phase" where known hotspots and non-hotspots patterns are fed to the system and "The detection phase" where hotspots are detected in testing layouts. In the training phase, (Figure 5.5), given the training layout clips, the known hotspot and non-hotspot patterns are first grouped

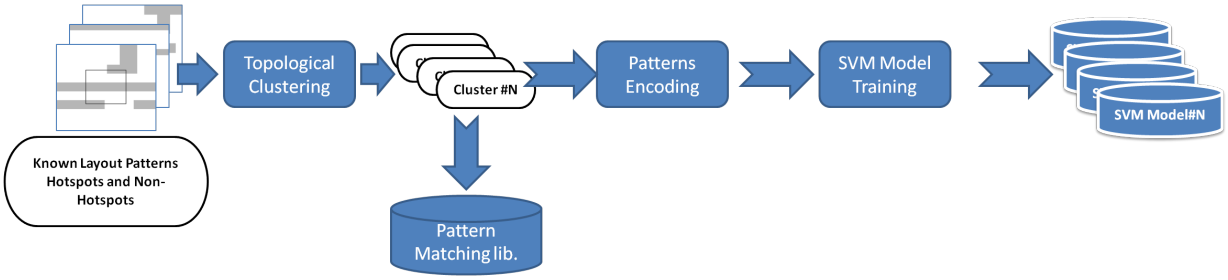


Figure 5.5: Training Phase

.

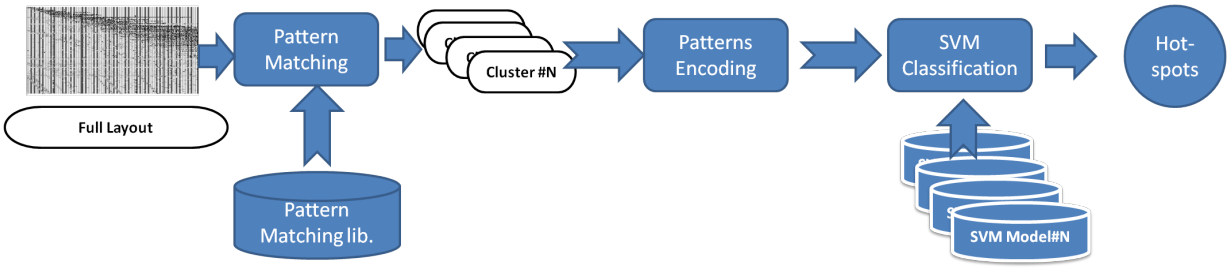


Figure 5.6: Detection Phase

.

into clusters according to their topologies. The key features of each cluster are stored in a pattern matching library. Secondly, for each cluster group a set of hotspot patterns and a set of non-hotspot are collected. Thirdly, critical features are extracted from each pattern and encoded into several fragment vectors. Finally, a specific SVM kernel is constructed for each cluster.

In the detection phase, (Figure 5.6), the full layout is initially scanned using pattern matching to tag the potentially problematic areas. Then, each location is classified into which cluster it belongs to. The pattern is then segmented into fragments and we calculate the encoded fragment vector for each fragment. Finally, the SVM classifier would then decide if the fragment in question is a hotspot or not.

Each cluster is independent from the others. This maintains a high flexibility in the system to add or remove clusters of patterns as the process technology is modified or more data is obtained from silicon measurement.

5.4.2 Topological Clustering

Hotspots can be very different. Some patterns may be simple 1D structures, and others can involve several 2D patterns. Attempting to train all data in a single general classifier would degrade the classification performance, because the training data becomes too complicated and the training time becomes time-consuming. To simplify the problem of building the classifier, we divide the hotspot detection problem to many simpler problems. We group the training data according to their topological information into different clusters. Each cluster will be treated as a separate classification problem, and the results of the combined classifiers will represent the entire systems results.

Figure 5.7 illustrate an example where four patterns are clusters into two different groups. The middle part of Figure 5.7 (a),(b) share a common topological pattern (A) and hence are grouped together. Similarly patterns in Figure 5.7 (c),(d) are grouped together in a cluster defined by the common topological pattern (B). In this work we used *Mentor Graphics's Calibre PatternMatching* [88] for clustering similar patterns. With topological clustering, each SVM kernel can concentrate on the critical features specific to its corresponding cluster, as well as provide a flexibility to identify previously unseen patterns. Topological clustering also facilitates hotspot and non-hotspot population balancing.

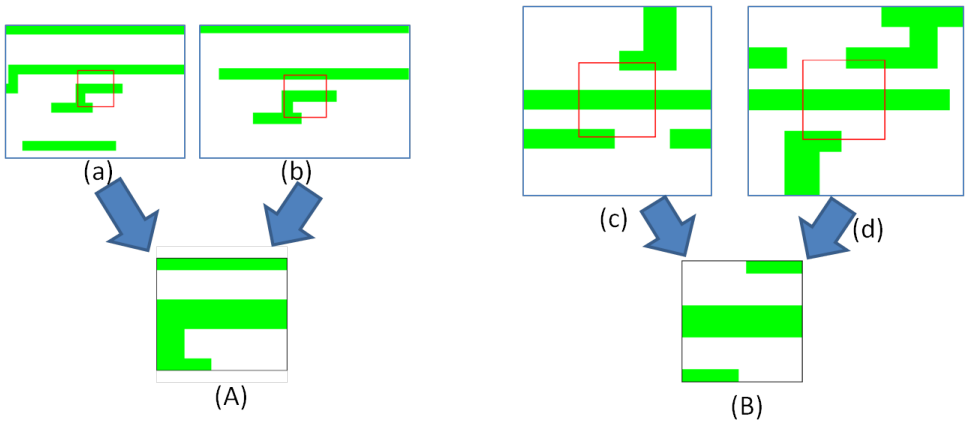


Figure 5.7: Pattern Clusters

.

5.4.3 Patterns Encoding

The first step in addressing the hotspot detection problem is to find a representation for the layout patterns that is sufficient to describe the layout environment causing the hotspots. Two dimensional irregularity in layouts is a major cause for hotspots [89], so in this work, we will combine the regularity metrics proposed in the previous chapter (section 4.4) with the concept of the fragmentation based context characterization [79] to encode the layout patterns. This encoding technique makes the problem of classification of hotspots easier because it describes the patterns based on key aspects that decide if the pattern is regular or irregular.

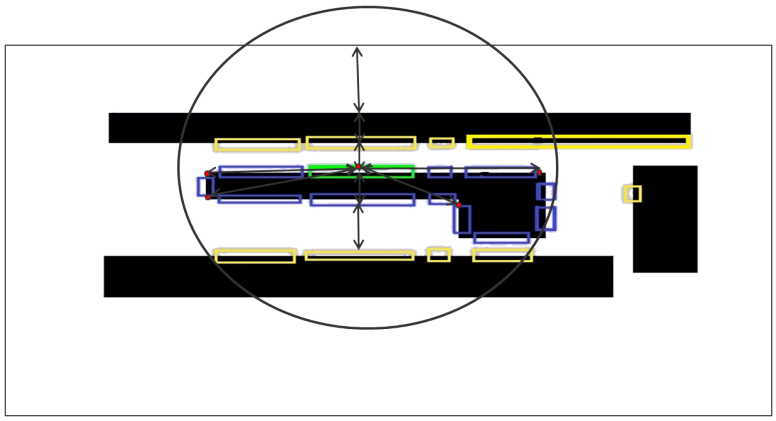


Figure 5.8: Regularity based Pattern Encoding

.

As shown in figure 5.8, the representation is based on 6 factors that will be presented as a vector of 9-dimensions:

1. External space,
2. Internal width,
3. Internal width for the externally facing fragment,
4. External space for the internally facing of the externally facing fragment,

5. External space for the internally facing fragment,

6. Distances to the nearest 4 corners.

This representation captures both the design pitch and the design two-dimensional irregularity as a measure for the pattern regularity. An advantage of the extracted encoding is that it is immune against symmetric mirroring, flipping and rotation of patterns.

5.4.4 Supervised Training System

Since SVM based systems showed better accuracy than ANN systems [79], we have chosen SVM as our method for training and classification of encoded patterns. Because of the nature of the problem, there are many patterns of various shapes in a layout, but only a few of these patterns are hotspots. The problem is that non-hotspot patterns greatly outnumbers hotspot patterns. Imbalanced data sets, where one class of data far outnumbers the other class, drop the accuracy of SVM significantly [90]. To enhance the accuracy of the training phase, a data sampling technique is proposed to balance the number of non-hotpot data. Since *Support vectors* are the data points that lie closest to the decision surface in SVM, and they are the most difficult to classify and have direct effect on the location of the optimized decision surface [87], it is important when sampling non-hotspot data to include these data points that are closest to the hotspot data. A brute-force algorithm is used to calculate the minimum Euclidean distance between each non-hotspot data point and the hotspot data points. Where the Euclidean distance $d(x, y)$ between two J-dimensional vectors x and y is defined in Eq.5.4

$$d(x, y) = \sqrt{\sum_{i=1}^{J} (x_i - y_i)^2}. \tag{5.4}$$

Only non-hotspot data points that are within a certain distance from the hotspot data points are included in the training.

5.5 Experimental Results

The algorithm was implemented in C++ programming language with the SVM library LIBSVM [87], integrated with Mentor Graphics Calibre tool for layout fragmentation and pattern matching [88]. Experiments were executed on a platform with four Intel Xeon 3.2 GHz CPUs and with 12 GB memory. To study the ability of the proposed flow to predict hotspots that were not in the training data set, we will compare its results against the results of pattern matching results, where we will use the hotspots in the training data set to define a library of patterns to be exactly matched in the testing layouts. Also to evaluate the effect of using topological clustering, we will examine a simpler flow with only single SVM kernel trained by the entire training data set. Results from our proposed flow will be compared against the single SVM kernel system to show the benefit of using topological clustering.

Table 5.1: ICCAD 2012 Benchmarks statistics

Training Data Set			Testing Layouts	
Name	HS#	NHS#	Name	HS#
Training1	99	340	B1	226
Training2	174	5285	B2	498
Training3	909	4643	B3	1808
Training4	95	4452	B4	177
Training5	26	2716	B5	41

The proposed approach is tested on the benchmarks released in [85]. Table 5.1 shows the statistics of five industrial benchmarks. The training layouts are composed of different clips that represent hotspot and non-hotspots clips. The number of hotspot (HS#) and non-hotspot (NHS#) clips are shown. For each training data set, a testing layout is provided to verify the accuracy of the hotspot detection flow. A true hotspot that is detected by the system will be reported as a *Hit*, while the hotspot that is not detected will be called a *Miss*. Falsely detected hotspot will be called an *Extra*. Two metrics are defined *Accuracy* and $Hit/Extra$ ratio to evaluate the performance of the hotspot identification methodology.

Table 5.2: Comparison of results with and without clustering

Benchmarks	PM Only	Single SVM		PM & ML	
	Hit	Hit	H/E	Hit	H/E
B1	44.7%	82.3%	0.38	73.4%	0.23
B2	34.7%	77.7%	0.10	83.1%	0.18
B3	50.9%	98.5%	0.04	99.7%	0.04
B4	54.3%	88.6%	0.05	97.7%	0.06
B5	63.4%	82.93%	0.07	85.4%	0.08
Average	49.6%	86%	0.13	87.9%	0.12

Both metrics should be maximized (5.5).

$$Accuracy(\%) = \frac{\#Hit}{\#HS}$$
$$H/E = \frac{\#Hit}{\#Extra} \tag{5.5}$$

Table 5.2 shows the results of the proposed flow. Accuracy of average 88% is achieved, and we also maintained a good Hit/Extra ratio. In the same table we compare the proposed flow with a pattern matching only solution. As expected, pattern matching fails to predict hotspots that were not part of the training set but on the other hand has zero false hits (Extras). This matches the findings in [81], where they reported that pattern matching has a low *Predictive Accuracy Rate* for unseen patterns but has a high *Memorizing Accuracy Rate* for seen patterns.

In the second set of experiments, as listed in Table 5.2, the effectiveness of topological clustering is demonstrated. Single SVM means the baseline SVM which uses one single huge SVM kernel (i.e., without topological classification). Except for the benchmark (B1), the use of clustering and an SVM kernel for each cluster of patterns significantly improved the accuracy of the prediction with maintaining (or slightly improving) the false hit rate. The low accuracy results of B1 is because the presence of new clusters of hotspots in the testing layout that were not part of the training set. This causes the pattern matching step to miss these patterns and hence are not input to the SVM classifiers.

Table 5.3: Runtime analysis

Operation / testcase	B1	B2	B3	B4	B5
Clustering + Encoding	7	320	5083	128	26
Train models	4	162	1879	108	4
Prediction	6	516	3253	261	37

The system training runtime and the prediction runtime are in the same order of magnitude. Table 5.3 shows the runtime of each step for the five testcases in the benchmark. The first row is the time taken to topologically cluster the training clips into different clusters and encode each clip using the encoding technique described in 5.4.3. The second row is the time taken to train the SVM models, the time varies for different testcases according to the number of training data points. Both rows one and two, represent the steps needed for system training. The third row is the time taken to run the prediction flow on the testing layouts.

Table 5.4: Comparison with other methods

Method	Accuracy	H/E	accuracy x (H/E)	CPU.hr/mm^2
Our results	87.86%	0.12	11	0.87
Ref. [80]	84.11%	0.10	8	0.87
Ref. [83]	92.7%	0.08	7	0.37
Ref. [86]	74.42%	2.53	188*	0.39

Table 5.4 shows our results compared with other reported methods on the same benchmark. The accuracy results come second after [83] that has the worst results in terms of H/E. Meanwhile our H/E results comes second after [86] that also has the worst results in terms of accuracy. In order to compare different methods, and since both accuracy and H/E metric are equally important, we drive a figure of merit that is the multiplication of both values. We will use this figure of merit to compare the different techniques. Except for [86] that has an accuracy less than 80%, which is the accepted level according to [85], our method shows the highest combined accuracy and H/E product. This shows that the method proposed is optimized in terms of both accuracy and H/E compared to the other

reported methods.

5.6 Conclusion

A hotspot detection system is demonstrated based on a hybrid pattern matching-SVM classifier. The integration of both pattern matching and machine learning techniques provides high accuracy and maintains the ability of the system to predict new hotspots. Patterns clustering and data balancing techniques are provided to enhance the performance of the proposed method. The system can adapt to changes in lithography process by only updating the classifiers that are related to the clusters affected by the process changes.

The experimental results show that the proposed approach effectively provides high accuracy in predicting unseen hotspots and minimizes false alarms. Comparing the results of the proposed approach to other published machine learning results showed that the proposed one is more effective than the other methods in providing high accuracy while simultaneously maintain a good false alarms rate.

Chapter 6

Localized Fixing of Catastrophic Hotspots in Interconnects

6.1 Introduction

Lithography hotspots are layout patterns that are sensitive to variations in the lithography processes and negatively affect manufacturing yield. Critical Hotspots found during the design stage should be fixed before releasing (taping-out) the design to the manufacturing foundries. In order to fix the lithographic hotspots, lithographic knowledge and experience are usually needed, which is not common in design teams. This makes it difficult to determine the optimum layout modification to fix the hotspots without introducing new hotspots. The design team usually needs to consult with the process team on what correction is needed and then implement the fix and re-qualify that the change is correct. This is usually an iterative process that is very time-consuming. This makes manual fix of hotspots too expensive. So to help designers fix hotspots, an automated hotspot fixing system is needed.

In this chapter, the development of an automated hotspot fixing system applicable to metal layers is reported. The proposed hotspot system attempts to improve the regularity of location around the hotspot by making localized changes in the design while maintaining the connectivity and ensuring the design rule correctness of the modifications. The rest of

the chapter is organized as follows. We start by a review of the state-of-the-art hotspot fixing techniques in Sec. 6.2. For completeness sake, the layout regularity metric derived in 4.4 is stated in Sec. 6.3. Then the fixing system based on the layout regularity metric is explained in Sec. 6.4. Finally, experimental results are reported and discussed in Sec. 6.5, followed by the conclusion in Sec. 6.6.

6.2 State-of-the-Art in Hotspots Fixing Techniques

In general, the problem of automatic fixing of hotspots has been addressed in two kinds of approach. The first one is through the use of the Place-and-Route (PnR) tools. The PnR tool is instructed to change the routing path causing the hotspot from one track to another. This approach is usually referred to as Rip-and-Reroute. The second approach is to perform localized changes in the wires, such as modifying the widths of some wires, moving wires or modifying the vias. This approach usually includes off-router-grid changes to the design. Rip-and-Reroute approach has the advantage that it can be easily integrated in the PnR tools and does not change the design flow, while having high risk of introducing new hotspots adjacent to the modified locations. Because Rip-and-Reroute approach may cause some routes to be modified dramatically, wire length and timing characteristics needs to be re-calculated to ensure minimum effect on circuit performance. The approach to localize the fixes with surgical off-grid changes guarantees that the timing characteristics of the design is intact.

6.2.1 Rip and Re-route fixing approach

This approach utilized router tools to fix hotspots. Hotspots are removed through iterations of ripping-up and rerouting paths one hotspot at a time. This iterative process can be time consuming because the new routes may cause new hotspots so each iteration of re-route requires a step of validation that usually includes expensive litho-simulaiton. Authors of [48,49] proposed guiding the re-routing by litho-simulations to reduce the number of iterations. Yang et al., [91] used Boolean Satisfiability (SAT) to simultaneously rip up

and re-route multiple nets in each hotspot region. This technique achieved fixing rate over 90% in only two iterations. The system is constrained by a set of pre-built library of known hotspot patterns that are forbidden to appear in the reroute. This limits the system to only known hotspots in the library, and cannot prevent the generation of new hotspots that are not captured in the library. The authors of [92] used an optical simulation engine to calculate fix guidance for each hotspot. The fix guidance is generated to optimize the optical intensity gain for edge movement near the hotspot. The fix guidance is then fed to the router to modify the layout and remove the hotspot.

6.2.2 Localized (Surgical) fixing approach

Rule-Based corrections that modify patterns in a pre-determined way were proposed by using local fixes [93], such as adding stubs or shifting jogs, to reduce the likelihood that fixing one hotspot creates another. In this proposal, the correction rules were tightly linked to detection rules in such a way that the proposed system does not detect a hotspot unless it knew how to correct it.
The hotspot fixer system (HSF) was proposed by [94, 95]. For each hotspot, modification rules are generated in terms of Line-Sizing and Space-Sizing, and are affecting the edges near the hotspot. The modification rules are then applied to the layout to generate a modified layout with fixing the hotspot. The modification rules can involve multiple patterns on multiple layers. To overcome limitations in the initial implementation of HSF, the authors of [96] improved the fixing rate by adding more rules to fix hotspots. The new rules include adding dummy and SRAF patterns in empty areas in the design near the hotspot, and also extend line ends to vacant areas if possible. Since the modification rules are blindly applied to all hotspots, the system generates 26 different candidate fix for each hotspot, and then runs litho-simulation to choose which fix actually removes the hotspot. Authors of [97, 98] proposed an approximate model-based repair hints flow. They first calibrate the approximate models using lithography simulator. For each category of structures, they develop a separate "Feature Model". This feature model predicts the impact of the design layout changes on the litho-contours. Then, for each litho-hotspot,

different topological features around the hotspot are extracted, and the effect of moving each edge is studied using the calibrated feature models and a collection of candidate edges movement are evaluated. Finally, hints are generated in a format that may be accepted by physical implementation tools.
The use of the PnR tool to perform surgical fixes around the hotspot was proposed in [99]. The surgical fixes such as cutting, padding, and moving some polygons to off-grid, were demonstrated and it was shown that with surgical fixes, no effect on timing was observed. The work in [100], used both approaches to fix hotspots. They started by rip-up and re-routing around hotspots, and then the newly introduced hotspots are fixed by localized guided-repair approach.

6.3 Layout Regularity Metric

The fact that variations sensitivity is pattern dependent raised the interest in studying the relation between the layout topology and variations sensitivity in different process steps. Since regularity leads to better control of process variability, industry started heading to more regular solutions for layout design. Examples of these are the work done in restrictive design rules (RDR) and regular fabrics. RDR is a post tape-out solution [101] while "regular fabrics" is a whole new design platform [102–104]. In previous chapter 4.4 and in [89], we reported a common definition for regularity that would decrease variations induced in each of the process steps including lithography, etching, rapid thermal annealing and chemical mechanical polishing [14, 105].

A metric was derived using a simple equation shown in Eq. 6.1. This equation contains geometrical properties such that the metric has a maximum value when the pattern resembles the most regular pattern. This means that the regular pattern has: (a) single orientation, (b) regular density, and (c) regular pitch. The metric value decreases as the pattern has line ends, jogs, corners and shapes of different orientations and different densities.

$$RM \propto \frac{\sum \text{lengths of edges in favored orientation}}{\sum \text{lengths of edges in unfavored orientation}} \times \frac{\sum \text{perimeter (shapes of layer)}}{\sum \text{area (shapes of layer)}} \times \frac{\sum \text{perimeter (shapes of derived layer}}{\sum \text{area (shapes of derived layer)}} \quad (6.1)$$

Where RM is the regularity metric.
"derived layer" is a layer created between the edges of projecting shapes within certain distance specified by the minimum spacing for each layer.
The regularity metric consists of three terms: the first term accounts for single orientation, the second term accounts for regular density and the third term accounts for regular pitch.

6.4 Proposed Flow

6.4.1 Overview

Figure 6.1: Hotspot fixing flow

The overall flow of the hotspot fix is shown in Fig. 6.1. For each detected hotspot, the engine will first find the edges that are expected to fix the hotspot if moved. Next, the system generates several clips with the candidate repair implemented and the candidate edge/edge-group is moved accordingly. Every clip is representing a possible fix for one litho-hotspot. Then, the fixes are sorted according to the regularity metric, such that the candidate fixes that are expected to improve the regularity are only processed. Then a validation step is run on the clip that has the highest regularity metric to validate that the hotspot is fixed and no new hotspots are introduced. If the fix is validated, it

```
procedure HotSpotFix(Design D , hotspot list hss)
    for all hotspot hs in hss do
        CandEdges ← FindCandidateEdges(hs)
        SortByDist(CandEdges)
        for all edge E in CandEdges do
            CandClips ← GenerateFixCands(E)
            SortByRegular(CandClips)
            for all candidate C in CandClips do
                hsNew ← ValidateClip(C)
                if hsNew = 0 then
                    fixCount ← fixCount
                    break
                end if
            end for
            if fixCount ≠ 0 then
                break
            end if
        end for
        ApplyFix(C)
    end for
end procedure
```

Figure 6.2: Hotspot Fix algorithm

will be applied to the design, and the next hotspot will be processed. For performance purpose, each hotspot can be processed in parallel on a separate thread. The hotspot fixing algorithm is shown in Fig. 6.2.

6.4.2 Repair Candidates

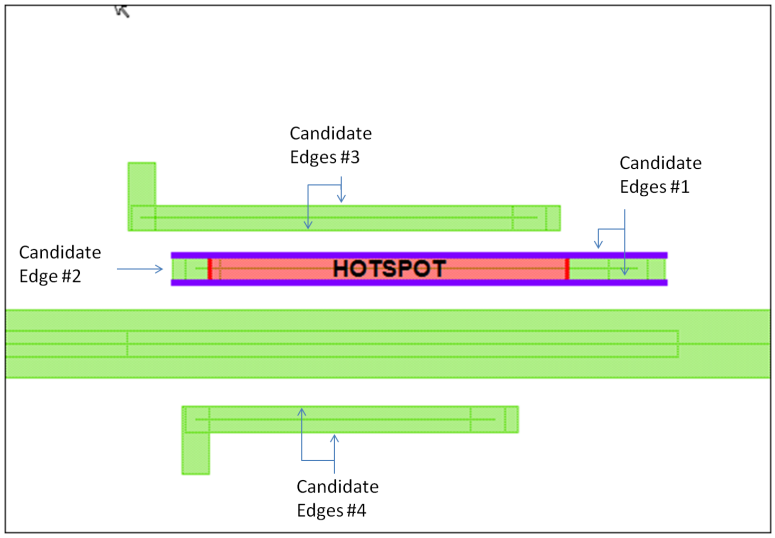


Figure 6.3: Fixing candidate edges.

The repair candidates clip generator module generates all the possible repair candidates for each hotspot. These repair candidates are edge-movement suggestions for edges in the proximity of the error marker. Movement suggestions can be in the form of single-edge movement or group-edge movement [106]. The system will first study the layout features in the proximity of the hotspot to identify a list of candidate edges/edge-groups that represent a possible cause of the hotspot as shown in Fig. 6.3. For each candidate in this list, a movement suggestion in the form of direction and value is generated. The generated candidates should not cause an internal or external DRC violation, i.e. applying those hints should not result in completely removing a polygon or making its width less than the minimum DRC width, and in the same time it should not result in merging two polygons or making the distance between them less than the minimum DRC space for this node.

With new technologies beyond the 20nm and introducing new manufacturing techniques such as double patterning (DP), it became a must to extend the DRC rules to also include different masks rules and include multiple layer DRC checks. The system depends on local optimizations around the hotspots like moving a single corner, a single line end or moving the hotspot's adjacent edges. These optimizations are sometimes restricted due to the strict design rules accompanying congested designs. A new algorithm is developed to identify a set of wire spreading candidates around the hotspot to be moved in case the single local optimization does not take place due to the restrictions above. This algorithm can move entire polygons instead of single or dual edges. Figure 6.4 shows a congested

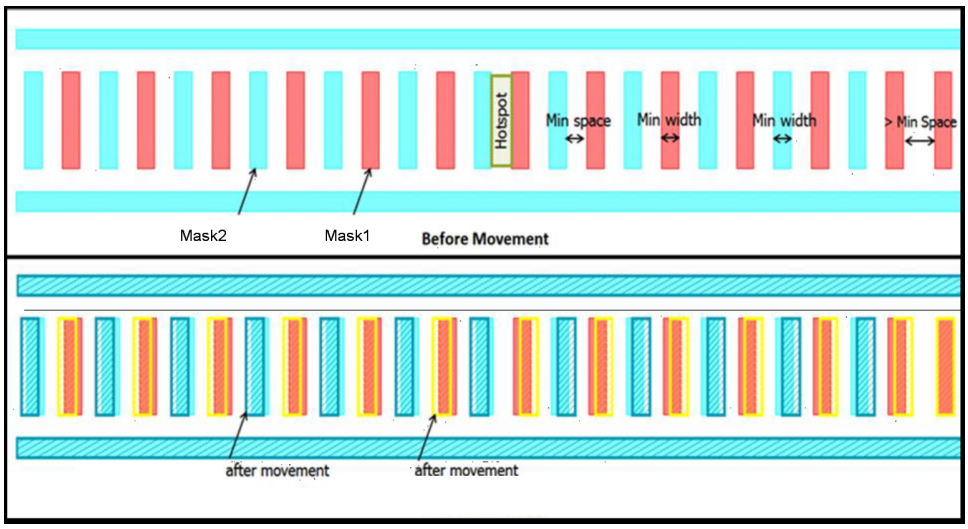


Figure 6.4: Wire spread example

DP design clip that has free spaces only near the edges of the clip, and the target layer is the DP metal layer. The generated candidate repair pushes the DP layer polygons on the left of the hotspot to the left and those on the right of the hotspot to the right, keeping it DRC clean. To preserve the design connectivity and also follow the DRC restrictions along multiple layers, the repair candidate generator supports multi-layer movement of edges/polygons.

If the movement of the candidate edge/polygon is restricted due to circuit connectivity violations, the blocking layer's edge/polygon is moved along with the set of wire spreading

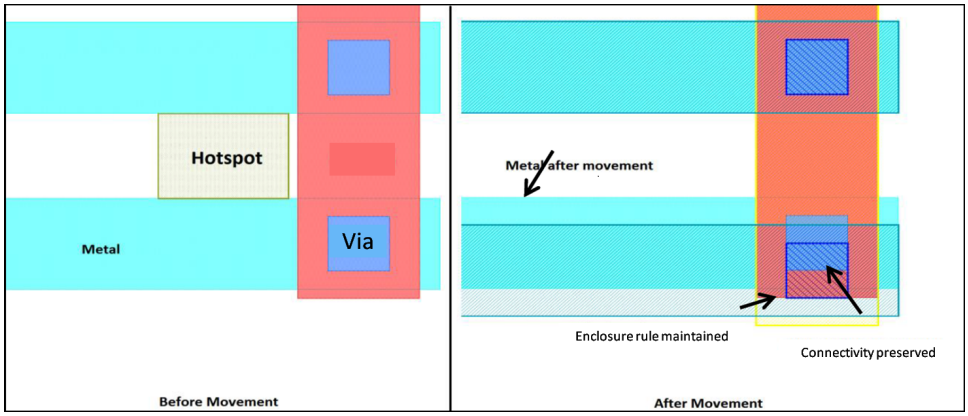


Figure 6.5: Multi-layer movement example

candidates created previously. Figure 6.5 shows a hotspot between two horizontal metal wires. The metal layer is the target layer for fixing. Moving the lower touching edge to the hotspot alone is forbidden, due to the connectivity specified between the horizontal metal and vertical metal, and moving the via along with the metal layer's edge alone will violate the enclosure rule specified between the via and vertical metal. The generated candidate solution moves 3 layers to maintain the connectivity and keep the design DRC clean.

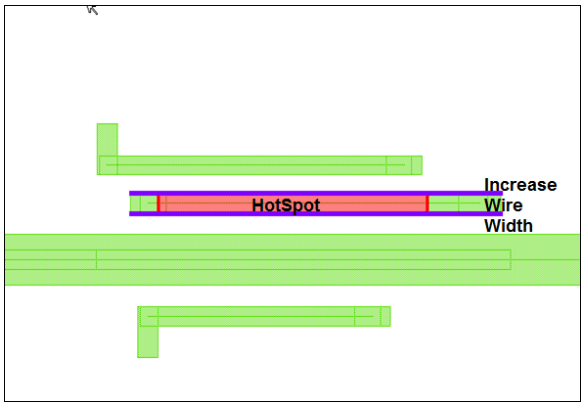
6.5 Experimental Results

The proposed flow for fixing hotspots was applied on a 32nm technology design. Lithography simulations found 83 hotspots in routing metal layers and 100 hotspots in metal1 layer.

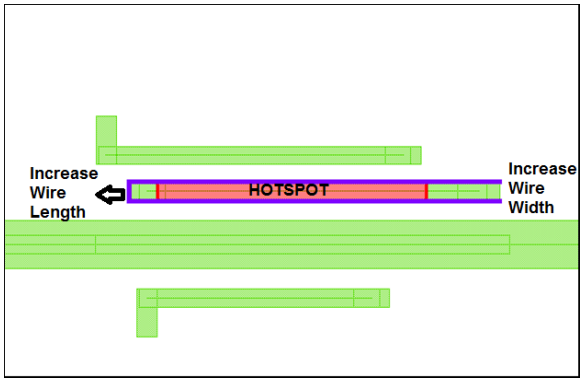
The repair candidates clip generator was used to generate five possible repair candidates for each hotspot of the 83 routing metal hotspots. The regularity metric Eq.6.1 was calculated for each repair candidate of each hotspot. It was observed that for each hotspot, some repair candidates improved regularity and others degraded the regularity.

Figures 6.6 and 6.7 show five repair candidates for the same hotspot, with their impact on the regularity metric. The candidate repairs can be to increase or decrease the width

of one or more wire or move certain wires. Candidate repairs can also be combination of multiple edge or wire movements. The lithography simulation confirms that the two repair candidates that improved the regularity metric in figure 6.6 actually fixed the hotspot, while the candidates that decreased the regularity metric in figure 6.7 did not fix the hotpot. This result matches the assumption that improving regularity is a good metric for judging if the repair candidate can actually fix the hotspot or not.



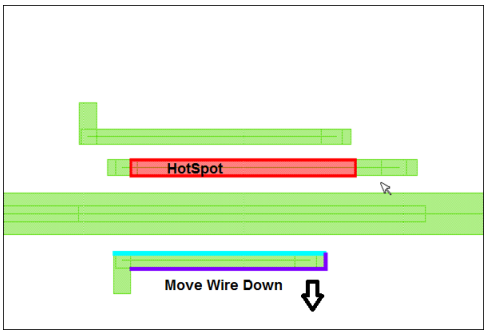
(a) Change in Regularity Metric 4.05%



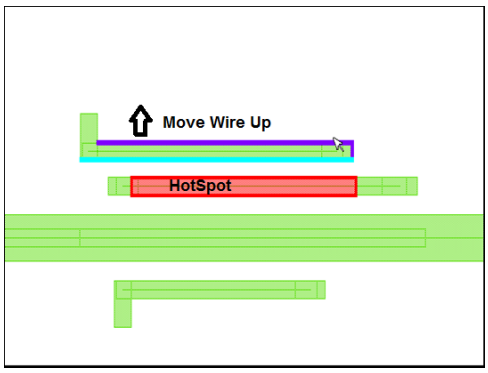
(b) Change in Regularity Metric 4.05%

Figure 6.6: Repair Candidates with improved regularity

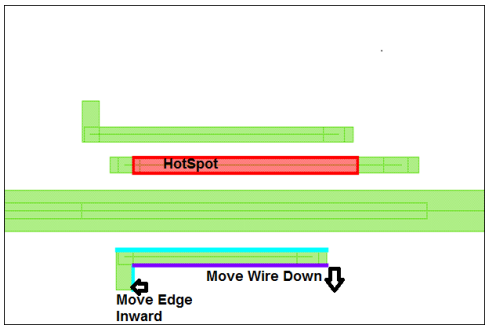
Out of the 83 Metal2 hotspots, the repair candidates generator was capable of finding one or more fixes for 82 hotspots, with a success rate 98.8% of fixing hotspots. 83% of the fixes are the candidate repair that maximized the regularity as defined by the metric



(a) Change in Regularity Metric -1.54%



(b) Change in Regularity Metric -1.75%



(c) Change in Regularity Metric -1.53%

Figure 6.7: Repair Candidates with decreased regularity

in Eq. 6.1, the remaining fixes aimed to relax the tight design constraint. Experiments were executed on a platform with four Intel Xeon 3.2 GHz CPUs and with 12 GB memory. The repair candidates generation step took less than 3 mins to generate the 415 repair candidates (five for each of the 83 hotspot). The regularity metric calculation for all the 415 repair candidates took less than 4 mins, resulting in total run time of 7 minutes with average runtime 5 seconds/hotspot. We repeated the same experiment but using lithography simulation to find the repair candidates that are validated to fix the hotspot. Runtime of the candidate generation and the lithography simulation phase combined was 74 minutes, with average runtime 53.5 seconds/hotspot.

We applied the same fixing methodology on the other 100 hotspots on Metal1. The repair candidates generator was capable of finding one or more fixes for all the hotspots,

with a success rate of 100% of fixing the hotspots. All of the fixes improved the regularity and in 97% of these cases the candidate with maximum value of regularity metric fixed the hotspot.

In Table 6.1, we compare the results obtained using our fixing system to other published results of various fixing techniques. The results show that our technique outperformed all the other reported results in terms of accuracy, while maintaining a comparable runtime of 5 seconds per hotspot. The reason for PnR based techniques [48, 49, 99] showing relatively low fixing rate, is there limitation to fixing routing metals and their inability to fix local metal1 hotspots. Pattern matching based technique [107] shows acceptable results with very fast runtime because of the nature of pattern matching, but is limited to only hotspots that are already calibrated in the pattern matching database and cannot predict or fix new hotspots that are not in the database. As this fixing technique is implemented in the place-and-route flow, it is limited to fixing the hotspots in the routing metals and cannot be extended to fix local metal1.

Table 6.1: Comparison Between various Fixing techniques

Technique	*Technology*	*HS Layers*	*HS count*	*Fix Rate*	*Runtime (sec/HS)*
our flow	**32nm**	**M1-M2**	**183**	**99%**	**5.0**
our flow + Simulation	**32nm**	**M1-M2**	**183**	**100%**	**53.5**
Wire Spreading + Rip-Reroute (RADAR) [48]	65nm	M1-M2	375	18%	2.2
Rip-Reroute (ELIAD) [49]	65nm	M1-M2	478	88%	1.6
Surgical Fix (HSF) [96]	28nm	not reported	120	81%	360
PnR basedSurgical Fix [99]	45nm	M1-M3	3961	41%	1.4
Pattern-Matching based Fix (PHR) [107]	32nm	M2-M4	19125	87%	0.5

6.6 Conclusion

In this chapter, a system for generating localized repair for fixing litho-hotspots was introduced. The generated fixes are satisfying two conditions. First, they do not violate the DRC constraints and preserve the connectivity along the layers stack. Second, they do not introduce new litho-hotspots. The system selects the repair candidate that maximizes the regularity. For a 32nm technology node design, it is shown that the success rate of the system in fixing hotspots is up to 97% and 83% for Metal1 and Metal2, respectively with runtime of 5 seconds/hotspot. Using lithography simulation to select the candidate repair increased the success of fixing rate to 100% and 98.8% for Metal1 and Metal2 respectively but increased the runtime more than an order of magnitude to be 53 seconds/hotspot.

Chapter 7

Conclusion and Future Directions

In this thesis we have presented a DFM framework for analyzing, mitigating and fixing the effect of process variation on circuit performance and manufacturing yield. We addressed the problem of process variations on both the device level (FEOL) and the interconnect level (BEOL). Table 7.1 summarizes the areas covered in this research in both the device and the interconnect layers.

7.1 Process Variation in the Device Level

Front-End-of-Line (FEOL) is the first portion of IC fabrication where the individual devices (transistors, capacitors, resistors, etc.) are manufactured in the semiconductor. This stage of manufacturing mostly suffers from parametric yield issues, where the variation in the process induces a large variation in circuit performance. Lithography variations and stress are the major contributors to the variation in the electrical performance of the devices. In chapter 4 we developed variation models for CMOS transistors that would convert the variations in the lithography demonstrated in the non-rectangular variations in the CMOS gate shape into variations in the transistor ON current and leakage current. We also integrated the effect of induced mechanical stress on the threshold voltage of the transistors.

Table 7.1: Summary of Work

	Device Layers (**FEOL**)	*Interconnect Layers* (**BEOL**)
Yield Limiting	Parametric Yield	Functional Yield
Process Variation	Litho+Stress	Litho
Effect	Timing Variation + Power Leakage	Open/Short Circuits
Analysis	Process Variation Models + Regularity Metric	Critical Features representation
Framework	Links Circuit to Physical	Machine Learning System
Mitigation/Fix	Regular Design	Automatic Fix flow
Design Stage	Std Cell + Logic Blocks	Post-Layout

In order to be able to use these transistor-level variation models in the analysis of digital circuits, we implemented an automated flow that bridges the gap between the digital circuit simulators, e.g. statistical timing analysis tools (STA), and the physical layout representation of the circuit. We have demonstrated the application of this flow using different digital benchmark designs using industrial 45nm technology. Using our flow and models, starting from the timing report generated from the STA tool, our flow was capable of identifying which transistors on the critical path that are most sensitive to process variations.

By analyzing the geometrical shapes that were the most sensitive to process variations, it was obvious that the geometrical irregularities that were making these shapes most sensitive to variations. This stimulated us to derive a simple regularity metric that would help us identify which shapes in the layout are irregular. In section 4.4, we developed such metric of quantitatively measuring regularity. We demonstrated that there is a strong correlation between irregularity and sensitivity to process variation, and it was sufficient to identify irregular shapes in the design to predict standard cells that will have high variability.

Starting from the 32nm technology node and beyond, to mitigate the effect of variations, the semiconductor industry and the design community have already established strong adherence to restricted design rules for FEOL layers. As shown in figure 4.14, Gate layer

has become very regular, with uniform one-dimensional pitch. As predicted by our study of the process variation models and the regularity metric, the adaption of this regular design style has decreased the negative effects of lithography and stress variations on the transistor variations. So although our flow, models and metric are applicable for newer technologies beyond 32nm, the need to perform such analysis is diminished thanks to the regular design style.

7.1.1 Future Directions in the variations of FEOL

In the context of modeling of variation of electrical parameters of devices, we will like to extend modeling to non-planner devices including 3D FinFET devices and nano-wires. We will also like to extend the modeling effects to include process variations from future EUV and Directed Self Assembly (DSA) technologies.

7.2 Process Variation in the Interconnects

The back-end-of-line (BEOL) is the second portion of IC fabrication where the individual devices get interconnected with wiring to create functioning circuits. We have focused on the lithographic yield limiting factors in this stage. Lithography variations cause wires to bridge or pinch and hence cause catastrophic open or short circuits, respectively. In chapter 5, we have developed a mechanism of detecting catastrophic failures in interconnects using machine learning approach. Using the same concept of regularity that was used in deriving the regularity metric in 4.4, we developed a pattern representation that would encode each pattern in a vector of 9 parameters. This representation based on regularity, would make the problem of classifying patterns into hotspots and non-hotspots easier.
We developed a flow using machine learning system, based on SVM classifiers to classify patterns. First, a supervised learning stage is required to train the SVM classifiers based on known hotspots and non-hotspots, then the system is used to predict hotspots in any layout. The accuracy of the system is then improved by utilizing data clustering and data sampling techniques. Using 28nm and 32nm benchmark designs, we have showed an

accuracy of 88% for detection of real hotspots. Compared to other published techniques on the same benchmark data, our method showed superior results in terms of both accuracy and false detection rate.
Finally, in chapter 6, we proposed an automated flow for fixing catastrophic failures in interconnects. The implemented flow targets to improve the regularity around the hotspot. The fixes are localized surgical changes in wire width or space. The reported fixes may also include small movement of wires, contacts or vias. The resulted fixes maintain the circuit connectivity, and do not violate the design rules. We have demonstrated a success fixing rate of 99% on 32nm industrial designs.

7.2.1 Future Extensions in variations of BEOL

Since our proposed techniques are extendible to multiple patterning technologies in 14nm and 10nm, we believe that our proposed methods for detecting and fixing hotspots in interconnects are applicable up to 10nm.
In the field of detection of hotspots, fast detection of hotspots resulting from direct self assembly (DSA) technology template variation is a challenge for this promising technology.
In the area of fixing critical failures, we will like to extend the algorithm of fixing to support specific design rule violations resulting from coloring conflicts in multi-patterning (triple/quadruple-patterning) technologies.

References

[1] G. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[2] G. Moore, "Excerpts from A Conversation with Gordon Moore: Moores Law." ftp://download.intel.com/museum/Moores_Law/Video-Transcripts/Excepts_A_Conversation_with_Gordon_Moore.pdf, 2005. [Online; accessed 19-Dec-2010].

[3] W. Shockley, "Problems related to pn junctions in silicon," *Solid-State Electronics*, vol. 2, no. 1, pp. 35–60, 1961.

[4] W. Schemmert and G. Zimmer, "Threshold-voltage sensitivity of ion-implanted mos transistors due to process variations," *Electronics Letters*, vol. 10, no. 9, p. 151, 1974.

[5] K. J. Kuhn, "Cmos transistor scaling past 32nm and implications on variation," in *IEEE journal of Advanced Semiconductor Manufacturing Conference (ASMC)*, pp. 241–246, 2010.

[6] S. Borkar, T. Karnik, and V. De, "Design and reliability challenges in nanometer technologies," in *Proceedings of the 41st annual Design Automation Conference*, p. 75, ACM, 2004.

[7] X. Lin and V. Moroz, "Layout Proximity Effects and Modeling Alternatives for IC Designs," *Design & Test of Computers, IEEE*, vol. 27, no. 2, pp. 18–25, 2010.

[8] S. Nassif, "Process variability at the 65nm node and beyond," in *Custom Integrated Circuits Conference*, pp. 1–8, IEEE, 2008.

[9] K. J. Kuhn, M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, "Process technology variation," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2197–2208, 2011.

[10] C. Mack, *Field guide to optical lithography*. SPIE Press, Bellingham, WA, 2006.

[11] W. Arnold, "Towards 3nm overlay and critical dimension uniformity: an integrated error budget for double patterning lithography," in *Proceedings of SPIE*, vol. 6924, p. 692404, 2008.

[12] I. Rangelow, "Critical tasks in high aspect ratio silicon dry etching for microelectromechanical systems," *Journal of vacuum science and technology. A. Vacuum, surfaces, and films*, vol. 21, no. 4, pp. 1550–1562, 2003.

[13] Y. Wei, J. Hu, F. Liu, and S. Sapatnekar, "Physical design techniques for optimizing RTA-induced variations," in *Design Automation Conference (ASP-DAC), 15th Asia and South Pacific*, pp. 745–750, IEEE, 2010.

[14] I. Ahsan, N. Zamdmer, O. Glushchenkov, R. Logan, E. Nowak, H. Kimura, J. Zimmerman, G. Berg, J. Herman, E. Maciejewski, *et al.*, "RTA-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology," in *Symposium on VLSI Technology,*, pp. 170–171, IEEE, 2006.

[15] Y. Ye, F. Liu, M. Chen, and Y. Cao, "Variability analysis under layout pattern-dependent rapid-thermal annealing process," in *Design Automation Conference, DAC. 46th ACM/IEEE*, pp. 551–556, IEEE, 2009.

[16] B. P. Wong, A. Mittal, G. W. Starr, F. Zach, V. Moroz, and A. Kahng, *Nano-CMOS Design for Manufacturability: Robust Circuit and Physical Design for Sub-65nm Technology Nodes*. New York, NY, USA: Wiley-Interscience, 2008.

[17] T. Chan, R. Ghaida, and P. Gupta, "Electrical Modeling of Lithographic Imperfections," in *2010 23rd International Conference on VLSI Design*, pp. 423–428, IEEE, 2010.

[18] X. Xie, *Physical understanding and modeling of chemical mechanical planarization in dielectric materials*. PhD thesis, Massachusetts Institute of Technology, 2007.

[19] S. Saha, "Modeling Process Variability in Scaled CMOS Technology," *Design & Test of Computers, IEEE*, vol. 27, no. 2, pp. 8–16, 2010.

[20] M. Abu-Rahma and M. Anis, "A statistical design-oriented delay variation model accounting for within-die variations," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 27, no. 11, pp. 1983–1995, 2008.

[21] C. Amin, N. Menezes, K. Killpack, F. Dartu, U. Choudhury, N. Hakim, and Y. Ismail, "Statistical static timing analysis: how simple can we get?," in *Design Automation Conference, DAC. Proceedings. 42nd*, pp. 652–657, IEEE, 2005.

[22] S. Nassif, "Modeling and forecasting of manufacturing variations (embedded tutorial)," in *Design Automation Conference (ASP-DAC), Asia and South Pacific*, pp. 145–150, ACM, 2001.

[23] E. Chin, C. Levy, and A. Neureuther, "Variability aware timing models at the standard cell level," in *Proceedings of SPIE*, vol. 7641, p. 76410, 2010.

[24] P. Gupta and F. Heng, "Toward a systematic-variation aware timing methodology," in *Proceedings of the 41st annual Design Automation Conference*, p. 326, ACM, 2004.

[25] D. Pan, "Lithography-aware physical design," in *ASIC. ASICON. 6th International Conference On*, vol. 1, pp. 1172–1173, IEEE, 2006.

[26] A. Kahng, C. Park, and X. Xu, "Fast dual-graph based hot-spot detection," in *Proc. 27th BACUS Symposium on Photomask Technology and Management*, pp. 6281–04, Citeseer, 2006.

[27] N. Rodriguez, L. Song, S. Shroff, K. Chen, T. Smith, and W. Luo, "Hotspot prevention using CMP model in design implementation flow," in *Quality Electronic Design. ISQED. 9th International Symposium on*, pp. 365–368, IEEE, 2008.

[28] P. Nag and W. Maly, "Hierarchical extraction of critical area for shorts in very large ICs," in *Defect and Fault Tolerance in VLSI Systems. International Workshop on,*, pp. 19–27, IEEE, 1995.

[29] T. Chan, R. Ghaida, and P. Gupta, "Electrical Modeling of Lithographic Imperfections," in *VLSI Design. VLSID. 23rd International Conference on*, pp. 423–428, IEEE, 2010.

[30] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proceedings of the 40th annual Design Automation Conference*, pp. 338–342, ACM, 2003.

[31] S. Nassif, "Variation in 45nm and implications for 32nm and beyond," in *NMI 2nd International Conference on CMOS Variability, ICCV*, NMI, 2009.

[32] M. Alam, K. Kang, B. Paul, and K. Roy, "Reliability-and Process-Variation Aware Design of VLSI Circuits," in *Physical and Failure Analysis of Integrated Circuits. IPFA. 14th International Symposium on the*, pp. 17–25, IEEE, 2007.

[33] W. Kuo and T. Kim, "An overview of manufacturing yield and reliability modeling for semiconductor products," *Proceedings of the IEEE*, vol. 87, no. 8, pp. 1329–1344, 2002.

[34] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits*, vol. 2. Prentice Hall, 2002.

[35] S. Nassif, "Design for variability in DSM technologies [deep submicron technologies]," in *Quality Electronic Design. ISQED. First International Symposium on*, pp. 451–454, IEEE, 2002.

[36] K. Bernstein, D. Frank, A. Gattiker, W. Haensch, B. Ji, S. Nassif, E. Nowak, D. Pearson, and N. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, no. 4/5, p. 433, 2006.

[37] M. Alioto, G. Palumbo, and M. Pennisi, "Understanding the Effect of Intradie Random Process Variations in Nanometer Domino Logic," *Integrated Circuit and System*

Design. Power and Timing Modeling, Optimization and Simulation, pp. 136–145, 2009.

[38] K. Kuhn, C. Kenyon, and A. Kornfeld, "Managing Process Variation in Intels 45nm CMOS Technology," *Intel Technology Journal*, vol. 12, no. 2, pp. 93 –109, 2008.

[39] A. Nardi and A. Sangiovanni-Vincentelli, "Synthesis for manufacturability: a sanity check," in *Design, Automation and Test in Europe Conference and Exhibition. DATE. Proceedings*, vol. 2, pp. 796–801, IEEE, 2004.

[40] H. Sunagawa, H. Terada, A. Tsuchiya, K. Kobayashi, and H. Onodera, "Effect of Regularity-Enhanced Layout on Variability and Circuit Performance of Standard Cells," *IPSJ Transactions on System LSI Design Methodology*, vol. 3, pp. 130–139, 2010.

[41] T. Jhaveri, V. Rovner, L. Liebmann, L. Pileggi, A. Strojwas, and J. Hibbeler, "Co-optimization of circuits, layout and lithography for predictive technology scaling beyond gratings," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 29, no. 4, pp. 509–527, 2010.

[42] K. Kuhn, "Variation in 45nm and Implications for 32nm and Beyond," in *2nd International CMOS Variability Conference*, 2009.

[43] P. Gupta, A. Kahng, Y. Kim, and D. Sylvester, "Self-Compensating Design for Reduction of Timing and Leakage Sensitivity to Systematic Pattern-Dependent Variation," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 9, pp. 1614–1624, 2007.

[44] A. Kahng, S. Muddu, and P. Sharma, "Detailed placement for leakage reduction using systematic through-pitch variation," in *Low Power Electronics and Design (ISLPED), ACM/IEEE International Symposium on*, pp. 110–115, IEEE, 2010.

[45] A. Kahng, C. Park, P. Sharma, and Q. Wang, "Lens aberration aware placement for timing yield," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 14, no. 1, pp. 1–26, 2009.

[46] S. Hu, P. Shah, and J. Hu, "Pattern Sensitive Placement Perturbation for Manufacturability," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 18, no. 6, pp. 1002 –1006, 2010.

[47] J. Yang, N. Rodriguez, O. Omedes, F. Gennari, Y. Lai, and V. Mankad, "DRCPlus in a router: automatic elimination of lithography hotspots using 2D pattern detection and correction," in *Proceedings of SPIE*, vol. 7641, p. 76410, 2010.

[48] J. Mitra, P. Yu, and D. Z. Pan, "Radar: Ret-aware detailed routing using fast lithography simulations," in *Design Automation Conference, DAC. Proceedings. 42nd*, pp. 369–372, IEEE, 2005.

[49] M. Cho, K. Yuan, Y. Ban, and D. Z. Pan, "Eliad: Efficient lithography aware detailed router with compact post-opc printability prediction," in *Proceedings of the 45th annual Design Automation Conference*, pp. 504–509, ACM, 2008.

[50] M. Cho, D. Pan, H. Xiang, and R. Puri, "Wire Density Driven Global Routing for CMP Variation and Timing," in *Computer-Aided Design. ICCAD. IEEE/ACM International Conference on*, pp. 487–492, IEEE, 2006.

[51] R. Ghaida and P. Gupta, "Design-overlay interactions in metal double patterning," in *Proceedings of SPIE*, vol. 7275, p. 727514, 2009.

[52] A. Balasinski, L. Karklin, and V. Axelrad, "Impact of subwavelength CD tolerance on device performance," in *Proceedings of SPIE*, vol. 4692, p. 361, 2002.

[53] R. Pack, V. Axelrad, A. Shibkov, V. Boksha, J. Huckabay, R. Salik, W. Staud, R. Wang, and W. Grobman, "Physical and timing verification of subwavelength-scale designs: I. Lithography impact on MOSFETs," in *Proceedings of SPIE*, vol. 5042, p. 51, 2003.

[54] 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," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 21, no. 5, pp. 544 –553, 2002.

[55] J. Yang, L. Capodieci, and D. Sylvester, "Advanced timing analysis based on post-OPC extraction of critical dimensions," in *Design Automation Conference, 2005. Proceedings. 42nd*, pp. 359–364, IEEE, 2005.

[56] P. Gupta, A. Kahng, S. Nakagawa, S. Shah, and P. Sharma, "Lithography simulation-based full-chip design analyses," in *Proceedings of SPIE*, vol. 6156, p. 61560, 2006.

[57] K. Cao, S. Dobre, and J. Hu, "Standard cell characterization considering lithography induced variations," in *Design Automation Conference. DAC. Proceedings. 43nd*, pp. 801–804, IEEE, 2006.

[58] D. James, "Design-for-manufacturing features in nanometer logic processes - a reverse engineering perspective," in *Custom Integrated Circuits Conference. CICC. IEEE*, pp. 207 –210, 2009.

[59] D. Sylvester, O. Nakagawa, and C. Hu, "Modeling the impact of back-end process variation on circuit performance," in *VLSI Technology, Systems, and Applications. International Symposium on*, pp. 58–61, IEEE, 2002.

[60] Y. Zhou, Z. Li, Y. Tian, W. Shi, and F. Liu, "A new methodology for interconnect parasitics extraction considering photo-lithography effects," in *Design Automation Conference. ASP-DAC. Asia and South Pacific*, pp. 450–455, IEEE, 2007.

[61] L. He, A. Kahng, K. H. Tam, and J. Xiong, "Simultaneous Buffer Insertion and Wire Sizing Considering Systematic CMP Variation and Random Leff Variation," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 5, pp. 845 –857, 2007.

[62] P. Gupta, A. Kahng, P. Sharma, and D. Sylvester, "Gate-length biasing for runtime-leakage control," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, no. 8, pp. 1475 – 1485, 2006.

[63] S. Teh, C. Heng, and A. Tay, "Design-process integration for performance-based OPC framework," in *Design Automation Conference. DAC. 45th ACM/IEEE*, pp. 522–527, IEEE, 2008.

[64] 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," in *Design Automation Conference. DAC. 45th ACM/IEEE*, pp. 516–521, IEEE, 2008.

[65] J. Wuu, F. Pikus, and M. Marek-Sadowska, "Fast and simple modeling of non-rectangular transistors," in *Proceedings of SPIE*, vol. 7122, p. 71223, 2008.

[66] C. Wang, W. Zhao, F. Liu, M. Chen, and Y. Cao, "Modeling of layout-dependent stress effect in CMOS design," in *Computer-Aided Design-Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, pp. 513–520, IEEE, 2009.

[67] A. Balasinski, "A methodology to analyze circuit impact of process-related mosfet geometry," in *Microlithography 2004*, pp. 85–92, International Society for Optics and Photonics, 2004.

[68] W. J. Poppe, L. Capodieci, J. Wu, and A. Neureuther, "From poly line to transistor: building bsim models for non-rectangular transistors," in *SPIE 31st International Symposium on Advanced Lithography*, pp. 61560–61560, International Society for Optics and Photonics, 2006.

[69] E. Shauly, A. Torres, L. Friedrich, M. Cohen-Yasour, O. Menadeva, and F. Pikus, "Full flow for transistor simulation based on edge-contour extraction and advanced SPICE simulation," in *Proceedings of SPIE*, vol. 7275, p. 727519, 2009.

[70] F. Pikus, "Integrated DFM framework for dynamic yield optimization," in *Proceedings of SPIE*, vol. 6349, p. 63490, 2006.

[71] T. Jhaveri, D. Motiani, K. Y. Tong, D. Pandini, T. Hersan, V. Kheterpal, V. Rovner, L. Pileggi, and A. J. Strojwas, "Maximization of layout printability/manufacturability by extreme layout regularity," *Journal of Micro/Nanolithography, MEMS, and MOEMS*, vol. 6, no. 3, pp. 031011–031011, 2007.

[72] M. Pons, M. Morgan, and C. Piguet, "Fixed origin corner square inspection layout regularity metric," in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 1397–1402, EDA Consortium, 2012.

[73] R. Aitken, G. Yeric, B. Cline, S. Sinha, L. Shifren, I. Iqbal, and V. Chandra, "Physical design and finfets," in *Proceedings of the 2014 on International Symposium on Physical Design*, ISPD '14, (New York, NY, USA), pp. 65–68, ACM, 2014.

[74] W. Hoppe, T. Roessler, and J. A. Torres, "Beyond rule-based physical verification," in *26th Annual BACUS Symposium on Photomask Technology*, pp. 63494X–63494X, International Society for Optics and Photonics, 2006.

[75] H. Yao, S. Sinha, C. Chiang, X. Hong, and Y. Cai, "Efficient process-hotspot detection using range pattern matching," in *Computer-Aided Design. ICCAD. IEEE/ACM International Conference on*, pp. 625–632, IEEE, 2006.

[76] J.-Y. Wuu, F. G. Pikus, A. Torres, and M. Marek-Sadowska, "Detecting context sensitive hot spots in standard cell libraries," in *SPIE Advanced Lithography*, pp. 727515–727515, International Society for Optics and Photonics, 2009.

[77] S. Mostafa, J. A. Torres, P. Rezk, and K. Madkour, "Multi-selection method for physical design verification applications," in *SPIE Advanced Lithography*, pp. 797407–797407, International Society for Optics and Photonics, 2011.

[78] J.-Y. Wuu, F. G. Pikus, A. Torres, and M. Marek-Sadowska, "Rapid layout pattern classification," in *Proceedings of the 16th Asia and South Pacific Design Automation Conference*, pp. 781–786, IEEE Press, 2011.

[79] D. Ding, A. J. Torres, F. G. Pikus, and D. Z. Pan, "High performance lithographic hotspot detection using hierarchically refined machine learning," in *Proceedings of the 16th Asia and South Pacific Design Automation Conference*, pp. 775–780, IEEE Press, 2011.

[80] J.-R. Gao, B. Yu, and D. Z. Pan, "Accurate lithography hotspot detection based on PCA-SVM classifier with hierarchical data clustering," in *SPIE Advanced Lithography*, pp. 90530–90530, International Society for Optics and Photonics, 2014.

[81] J.-Y. Wuu, F. G. Pikus, and M. Marek-Sadowska, "Efficient approach to early detection of lithographic hotspots using machine learning systems and pattern matching,"

in *SPIE Advanced Lithography*, pp. 79740–79740, International Society for Optics and Photonics, 2011.

[82] D. Ding, X. Wu, J. Ghosh, and D. Z. Pan, "Machine learning based lithographic hotspot detection with critical-feature extraction and classification," in *IC Design and Technology, 2009. ICICDT'09. IEEE International Conference on*, pp. 219–222, IEEE, 2009.

[83] Y.-T. Yu, G.-H. Lin, I. H.-R. Jiang, and C. Chiang, "Machine-learning-based hotspot detection using topological classification and critical feature extraction," in *Proceedings of the 50th Annual Design Automation Conference*, p. 67, ACM, 2013.

[84] D. Ding, B. Yu, J. Ghosh, and D. Pan, "EPIC: Efficient prediction of IC manufacturing hotspots with a unified meta-classification formulation," in *Design Automation Conference (ASP-DAC), 2012 17th Asia and South Pacific*, pp. 263–270, Jan 2012.

[85] J. A. Torres, "ICCAD-2012 CAD Contest in Fuzzy Pattern Matching for Physical Verification and Benchmark Suite," in *Proceedings of the International Conference on Computer-Aided Design*, ICCAD '12, (New York, NY, USA), pp. 349–350, ACM, 2012.

[86] S.-Y. Lin, J.-Y. Chen, J.-C. Li, W.-y. Wen, and S.-C. Chang, "A Novel Fuzzy Matching Model for Lithography Hotspot Detection," in *Proceedings of the 50th Annual Design Automation Conference*, DAC '13, (New York, NY, USA), pp. 68:1–68:6, ACM, 2013.

[87] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at urlhttp://www.csie.ntu.edu.tw/ cjlin/libsvm.

[88] Mentor Graphics, *Calibre Pattern Matching v2014.2*. Wilsonville, OR, 2014.

[89] E. Swillam, K. Madkour, and M. Anis, "Layout regularity metric as a fast indicator of high variability circuits," in *SOC Conference (SOCC), 2013 IEEE 26th International*, pp. 43–48, IEEE, 2013.

[90] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *In ICML Workshop on Learning from Imbalanced Data Sets*, pp. 49–56, 2003.

[91] F. Yang, Y. Cai, Q. Zhou, and J. Hu, "Sat based multi-net rip-up-and-reroute for manufacturing hotspot removal," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010*, pp. 1369–1372, IEEE, 2010.

[92] Y.-S. Tong and S.-J. Chen, "An automatic optical simulation-based lithography hotspot fix flow for post-route optimization," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 29, no. 5, pp. 671–684, 2010.

[93] G. T. Luk-Pat, A. Miloslavsky, A. Ikeuchi, H. Suzuki, S. Kyoh, K. Izuha, F. Tseng, and L. Wen, "Correcting lithography hot spots during physical-design implementation," in *Proc. SPIE*, vol. 6349, p. 634920, 2006.

[94] T. Kotani, S. Kyoh, S. Kobayashi, T. Inazu, A. Ikeuchi, Y. Urakawa, S. Inoue, E. Morita, S. Klaver, T. Horiuchi, *et al.*, "Development of hot spot fixer (hsf)," in *SPIE 31st International Symposium on Advanced Lithography*, pp. 61560–61560, International Society for Optics and Photonics, 2006.

[95] S. Kobayashi, S. Kyoh, T. Kotani, S. Tanaka, and S. Inoue, "Automated hot-spot fixing system applied to the metal layers of 65-nm logic devices," *Journal of Micro/Nanolithography, MEMS, and MOEMS*, vol. 6, no. 3, pp. 031010–031010, 2007.

[96] M. Kajiwara, S. Kobayashi, H. Mashita, R. Aburada, N. Furuta, and T. Kotani, "Configurable hot spot fixing system," in *SPIE Advanced Lithography*, pp. 90530–90530, International Society for Optics and Photonics, 2014.

[97] M. Chew, T. Endo, and Y. Yang, "prsm: models for model-based litho-hotspot repairs," in *SPIE Photomask Technology*, pp. 74883–74883, International Society for Optics and Photonics, 2009.

[98] Y. Yang, M. Chew, T. Endo, and M. Simmons, "Model-based hints for litho-hotspots repair," in *SPIE Photomask Technology*, pp. 74883–74883, International Society for Optics and Photonics, 2009.

[99] J. H. Park, S. W. Paek, N. Ha, D. H. Jang, B. M. Kim, H. S. Won, and K. M. Choi, "Fixing lithography hotspots on routing without timing discrepancy," in *SoC Design Conference (ISOCC), 2009 International*, pp. 33–36, IEEE, 2009.

[100] R. März, K. Peter, and K. Engelhardt, "Rerouting and guided-repair strategies to resolve lithography hotspots," in *SPIE Advanced Lithography*, pp. 79740–79740, International Society for Optics and Photonics, 2011.

[101] M. Lavin, F.-L. Heng, and G. Northrop, "Backend cad flows for restrictive design rules," in *Proceedings of the 2004 IEEE/ACM International conference on Computer-aided design*, pp. 739–746, IEEE Computer Society, 2004.

[102] L. Liebmann, L. Pileggi, J. Hibbeler, V. Rovner, T. Jhaveri, and G. Northrop, "Simplify to survive: prescriptive layouts ensure profitable scaling to 32nm and beyond," in *SPIE Advanced Lithography*, pp. 72750–72750, International Society for Optics and Photonics, 2009.

[103] B. Taylor and L. Pileggi, "Exact combinatorial optimization methods for physical design of regular logic bricks," in *Proceedings of the 44th annual Design Automation Conference*, pp. 344–349, ACM, 2007.

[104] D. Morris, V. Rovner, L. Pileggi, A. Strojwas, and K. Vaidyanathan, "Enabling application-specific integrated circuits on limited pattern constructs," in *VLSI Technology (VLSIT), 2010 Symposium on*, pp. 139–140, IEEE, 2010.

[105] K. O. Abrokwah, P. Chidambaram, and D. S. Boning, "Pattern based prediction for plasma etch," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 20, no. 2, pp. 77–86, 2007.

[106] A. Rabie, K. Madkour, K. George, W. ElManhawy, J.-M. Brunet, and J. Kwan, "Model based multilayers fix for litho hotspots beyond 20nm node," in *SPIE Advanced Lithography*, pp. 90530–90530, International Society for Optics and Photonics, 2014.

[107] D. Jang, N. Ha, J. Jeon, J.-H. Kang, S. W. Paek, H. Choi, K. S. Kim, Y.-C. Lai, P. Hurat, and W. Luo, "In-design process hotspot repair using pattern matching," in *SPIE Advanced Lithography*, pp. 83270–83270, International Society for Optics and Photonics, 2012.