Spatial Modeling of Gate Length Variation for Process-Design Co-Optimization

Paul David Friedberg

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-157
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-157.html

December 17, 2007

Spatial Modeling of Gate Length Variation for Process-Design Co-Optimization

by

Paul David Friedberg

B.A. (Williams College) 2001

M.S. (University of California, Berkeley) 2003

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering — Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Costas J. Spanos, Chair

Professor Jan Rabaey

Professor David Brillinger

Fall 2007

The dissertation of Paul David Friedberg is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Fall 2007

Abstract

Spatial Modeling of Gate Length Variation for

Process-Design Co-Optimization

by

Paul David Friedberg

Doctor of Philosophy in Engineering — Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Costas J. Spanos, Chair

Variability in circuit performance is a rapidly growing concern in the semiconductor industry, and a potential roadblock in circuit design. In order to combat the negative impact of manufacturing variations on circuit performance, two approaches can be taken. The first approach emphasizes process control from a manufacturing perspective, in an effort to directly reduce the variation in device parameters; the second approach attacks the problem from the design perspective, where specific design methodologies can be developed to decrease a circuit's sensitivity to process variation. Both approaches rely on exploration through the use of simulation frameworks that capture the detailed interaction between manufacturing variation and the resulting circuit performance variability. With such a framework, one can determine the most deleterious sources of device parameter variation, and then identify the effects of a certain flavor of process control, or search for sensitivity-reducing design techniques.

In order for the simulation framework to be truly useful, however, an accurate model of the relationship between variation and variability is required. This work explains how

a rigorous spatial statistical description of the manufacturing variation is a crucial aspect for the simulation framework. Using an array of test structures, spatial variation in critical dimension is exhaustively characterized. Then, a variety of statistical descriptions of spatial CD variation are instantiated in an analytical, macromodel-based Monte Carlo simulation framework. Based on evaluation of these statistical descriptions against one another, it is shown that a full decomposition of deterministic variation is required for optimal accuracy. Moreover, it is shown that such a decomposition of variance accounts for virtually all spatial autocorrelation in CD. Finally, it is shown that employing a simplified statistical description (as is commonly done in existing MC frameworks) that relies on spatial autocorrelation to capture deterministic variation overestimates the impact of variation on performance variability.

_____

Professor Costas J. Spanos
Dissertation Committee Chair

List of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The aggressive scaling of silicon technology has enabled dramatic improvements in integrated circuit performance [1.1]. However, as nominal device parameter values are rapidly reduced, control of semiconductor manufacturing processes (particularly lithography) has become increasingly difficult. For example, the 2003 edition of the International Technology Roadmap for Semiconductors [1.2] listed the control of printed transistor gate length in the lithography process as having an interim solution for 2003 but no known solution for the then-current 2004 technology generation. Similarly, the 2005 ITRS edition [1.3] listed gate length control as having a known interim solution for 2005, but no known solution for the 2006 generation. Based on these sources, it is clear that the semiconductor industry is barely meeting its needs in critical dimension (CD) control. As a result, variability in circuit performance is a rapidly growing concern in the semiconductor industry, and a potential roadblock in circuit design [1.4,1.5].

In order to combat the growing negative impact of manufacturing variations on circuit performance, two approaches are being taken. The first approach emphasizes process control from a manufacturing perspective, in an effort to directly reduce the variation in device parameters; in the lithography process, for example, developments of integrated, scatterometry-based metrology [1.6] and advanced post-exposure bake temperature uniformity control promise to improve gate CD control. The second approach comes from the design perspective, where specific design methodologies can be developed to decrease a circuit's sensitivity to process variation. Both approaches rely on exploration through the use of Monte Carlo simulation frameworks that can capture the detailed interaction between manufacturing variation and the resulting circuit performance variability. With such a framework, one may identify the most deleterious sources of device parameter variation and then characterize the effects of a certain flavor of process control, or search for sensitivity-reducing design techniques.

Traditionally, Monte Carlo simulation frameworks have been based upon rudimentary statistical models of manufacturing variation. Typically, these models are based upon the assumption that the variation in a given parameter can be fairly represented using a normal distribution; then, the framework designer simply assigns some mean and (often arbitrary) level of variance to each device parameter of interest. For each simulation run, values for each device parameter are drawn from the corresponding distribution in either (a) an entirely random fashion, with different values being chosen for each device in the given circuit, or (b) with perfect correlation—that is, each device in the circuit of interest is assigned the same randomly drawn device parameter value. However, as process control requirements become more difficult to achieve and robust design in turn becomes

a more difficult task, such rudimentary models fall short of the required statistical accuracy of a Monte Carlo. In order for the simulation framework to be truly useful, however, an extremely accurate model of the relationship between variation and variability is needed.

This work explains in detail why a rigorous spatial statistical description of manufacturing variation is such a critical component of Monte Carlo simulation frameworks. Using an array of test structures, spatial variation in critical dimension is exhaustively characterized. Then, a variety of statistical descriptions of spatial CD variation are instantiated in an analytical, macromodel-based Monte Carlo simulation framework. Based on evaluation of these statistical descriptions against one another, it is shown that a full decomposition of deterministic variation is required for optimal accuracy. Moreover, it is shown that such a decomposition of variance accounts for virtually all spatial autocorrelation in CD. Finally, it is shown that employing a simplified statistical description (as is commonly done in existing MC frameworks) that relies on spatial autocorrelation to capture deterministic variation overestimates the impact of variation on performance variability.

1.2 Dissertation Layout

Chapter 2 provides some basic introduction material. This includes a brief overview of the general semiconductor process highlighting potential sources of variation, as well as the associated implications of process variation on circuit variability. Current approaches for designing robust circuits will also be illustrated.

Chapter 3 contains a complete characterization of spatial variation of critical dimensions. First, the principal causes of CD variation as well as resulting process control opportunities are reviewed. Next, a historical gate length variation study aimed at exploring long range ($1mm$—$2cm$) variation is described. Finally, novel electrical linewidth metrology test structures for measuring short-range ($0.2\mu m$—$1mm$) spatial variation in gate length are presented, with a description of design, characterization, and results.

Chapter 4 details the construction of Monte Carlo simulation frameworks. Two different frameworks will be discussed: both a traditional SPICE-based Monte Carlo framework as well as an analytical, macromodel-based framework. Considerations of accuracy and execution time will be assessed.

In Chapter 5, the thorough characterization of CD variation from Chapter 3 is instantiated in the Monte Carlo simulation frameworks of Chapter 4 to enable exploration of the impact of CD variation on circuit performance variability. The importance of accuracy and completeness in the statistical description of process variation will be highlighted.

Chapter 6 is devoted to ongoing efforts to characterize additional sources of manufacturing variation. This includes the design of new test structures that interleave the novel ELM structures of Chapter 3 with an array of active MOS devices. Anticipated results will be shared.

Finally, a summary of the entire body of work is presented in Chapter 7. Ongoing and future directions for this work are addressed.

References

[1.1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics* **38**, April 19, 1965.

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

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

[1.4] B. E. Stine, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. **10**, no. **1**, Feb. 1997.

[1.5] S. Nassif, "Design for manufacturability in DSM technologies," *IEEE International Symposium on Quality Electronics Design*, pp. 451-454, 2000.

[1.6] X. Niu, N. Jakatdar, J. Bao, and C. Spanos, "Specular Spectroscopic Scatterometry," *IEEE Transactions on Semiconductor Manufacturing*, vol. **14**, no. **2**, pp. 97-111, 2001.

Chapter 2

Background

This chapter provides a basic introduction to semiconductor processing. An emphasis is placed on potential sources of variation as well as the associated implications of process variation on circuit variability. Current approaches for designing robust circuits will also be highlighted. The material in this chapter is not meant to be exhaustive, but rather to give the reader a sufficient level of understanding needed for the remaining chapters in this dissertation. For additional, more detailed explanation of concepts presented here briefly, the reader is advised to consult the included references as needed.

2.1 Introduction to the Pattern Transfer Process

The general flow for fabrication of integrated circuits is composed of several steps. In the front-end processes, dopants are implanted and diffused into the silicon substrate and various materials are repeatedly deposited and patterned to build active devices such as MOS transistors. In back-end processes, layers of interconnect used as wires between active devices are created using successive repetitions of deposition, patterning, and

polishing. This dissertation focuses primarily on variations that arise during the front-end pattern transfer processes, namely photolithography and etch. Therefore, a brief introduction will be provided to this segment of the general semiconductor process flow.

2.2 Photolithography

Photolithography is the first step in transferring a desired pattern into a thin film. Pattern transfer is achieved by using light exposure to change the chemical properties of a photosensitive material, called photoresist, which has been deposited on top of the device layer (thin film) to be patterned. By changing the chemical properties of the photoresist, it becomes either more resistant (in the case of positive resist) or less resistant (in the case of negative resist) to a particular solvent called a developer solution. With careful design, the exposed sections of a positive photoresist will have a pattern identical to that desired in the underlying device layer (or a negative image of the desired pattern in the case of negative resist), so that after the photoresist layer is placed in the developer, the pattern intended to be transferred into the device layer will remain in the form of photoresist. Finally, the underlying regions which are not protected by photoresist are etched away; when the remaining photoresist is stripped away, the layer has been patterned as desired.

2.2.1 Exposure

The most critical, difficult, and expensive step in the general lithography flow (shown in **Fig 2.1**) is the exposure step. The exposure system consists of three main components. First, a source of light at a particular wavelength is needed. Second, a reticle or mask is

used as a master copy of the desired pattern, through which the light is passed to transfer that pattern onto the photoresist-coated wafer. The current industry standard for the exposure step is projection printing, wherein the mask pattern is demagnified as it is transferred to the wafer. To accomplish this image reduction, the third component of the exposure system, a lens consisting of several optical elements (called projection optics), is required to project a reduced image of the mask onto the target wafer.

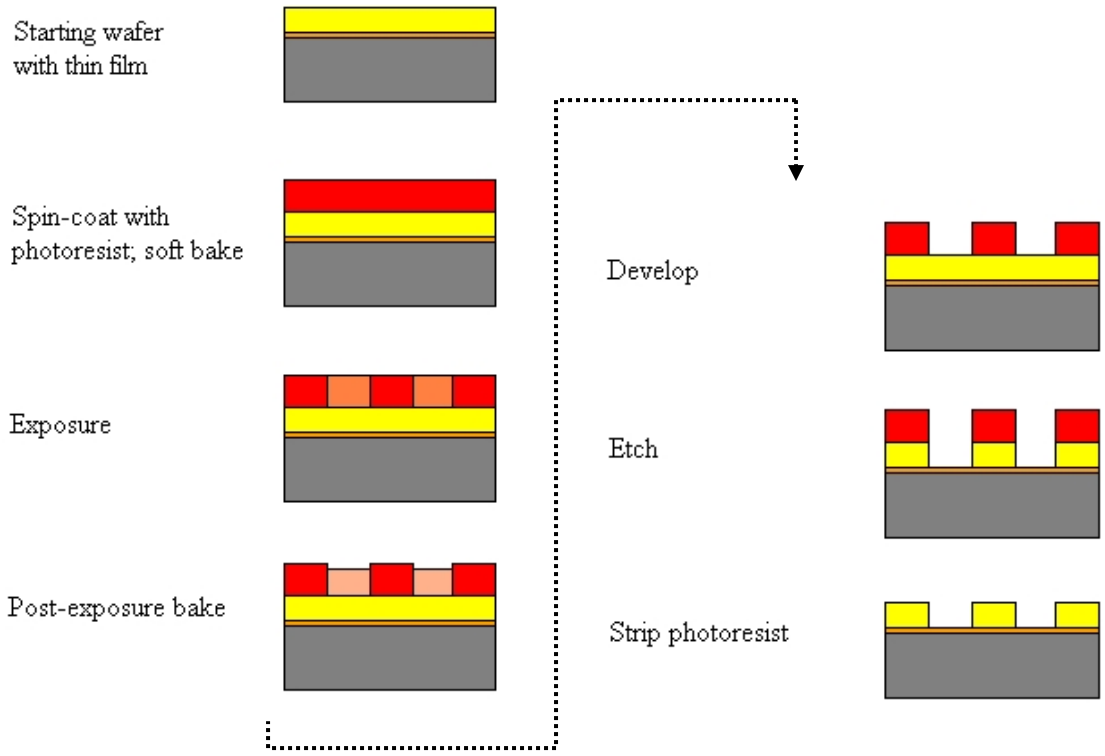

Figure 2.1. Illustration of the main steps in the lithography process flow. The etch step—the final step in the pattern transfer process—does not technically lie within the lithography module; however, it is included in this figure for completeness.

Projection optics give rise to diffraction effects, which in turn limit the minimum printable feature size of the lithography system, which can be expressed by the equation:

$$w_{\min} = k_1 \frac{\lambda}{NA} \tag{2.1}$$

where w_{min} is the minimum feature size, λ is the illumination wavelength, NA is the numerical aperture of the final lens element, and k_1 is a lithography "difficulty" factor, which effectively captures information about every other aspect of the lithography system. This equation is similar to the well-known Rayleigh criterion:

$$w_{\min} = 0.61\frac{\lambda}{NA} \tag{2.2}$$

where w_{min} corresponds to the minimum spacing between two points of light such that the intensity falls to 80% of its maximum value between the two points—a subjective criterion of minimum discernability. As we will see, however, in photolithography k_1 may be pushed well below Rayleigh's value of 0.61.

In order to improve the resolution of a lithography system, perhaps the most obvious solution is to reduce the illumination wavelength. The industry has shifted the illumination wavelength from the mercury e-line (577 nm) to the g-line (436 nm) and i-line (365 nm), and on to deep ultraviolet (DUV) wavelengths 248 nm (KrF excimer laser illumination), and, presently, to 193 nm (ArF excimer laser). Plans to eventually move to 157 nm (F2 laser) may not be realized, since most industry experts predict favor a direct transition to a lithography system that will use extreme ultraviolet (EUV) light, at a wavelength of 13 nm. However, the improvement in lithographic resolution has greatly out-stripped the reduction of illumination wavelength. At the 248-nm illumination wavelength node, lithographers began printing feature sizes that were smaller than the illuminating wavelength. This improvement has been possible due to the other two parameters in Equation 2.1, namely NA and k1. The numerical aperture is given by

$$NA = n(\sin\alpha)\,, \tag{2.3}$$

where α is one-half the angle of acceptance of the objective lens (the final lens element, which projects the mask image onto the wafer) and n is the refractive index of the medium separating the objective and the wafer. The step-wise exposure of single, small portions of the wafer (called fields) allows systems to be built with very high NAs, since the acceptance angle is greatly increased. Modern "dry" systems (where the medium between the wafer and the lens is air with $n \cong 1$) commonly have NAs over 0.8, compared to an NA around 0.25 in the 1970s, allowing for a three-fold improvement in resolution at a given wavelength.

As lens size increases, imperfections in the lens—or aberrations—become more difficult to avoid during lens manufacturing. To combat this difficulty, modern lithography systems limit the portion of the lens that is actually used during imaging. This is accomplished by using a slit centered over the lens that allows only a portion of the mask pattern to be imaged onto the wafer at one time. To print the entire field, the mask is then translated in one direction while the wafer is translated in the opposite direction in a scanning fashion (**Fig. 2.2**).

There are limits to the potential continued increase in NA. First, when imaging in air (as all current lithography systems do), NA cannot exceed $n_{air}(\sin\alpha) \approx 1$, since the index of refraction for air is approximately 1 and the sin(α) term cannot exceed 1. State-of-the-art immersion lithography systems increase NA to values greater than 1 (~1.25) through the use of a medium of higher refractive index (typically water; perhaps a perfluoropolyether liquid in the future) between the lens and wafer [2.1].

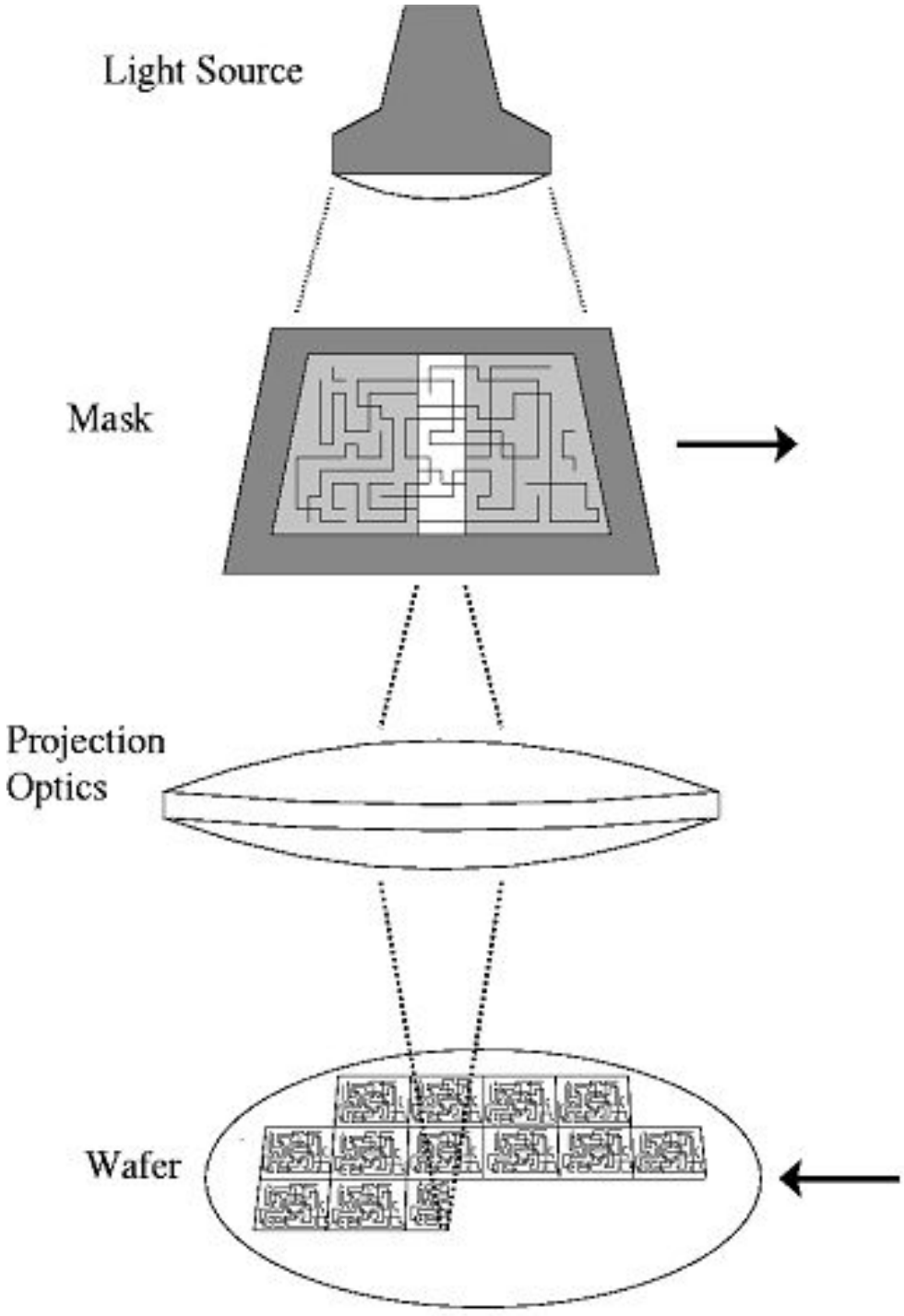

Figure 2.2. Illustration of the projection scanner exposure system. To print each lithographic field, the wafer and mask are translated in opposite directions in a scanning fashion.

However, the increase of NA has limitations, since the depth of focus—defined to be the range over which the wafer can be moved along the optical axis such that the image stays in focus—is inversely proportional to the square of NA:

$$DOF = k_2 \frac{\lambda}{NA^2} \tag{2.4}$$

If NA is increased too drastically, lithography will fail for even modest levels of topography on the wafer, since some features will fall outside of the usable focal range. Therefore, increasing NA has both physical limitations and additional processing constraints. Fortunately, NA increase is not expected to be the only vehicle for further improvement of lithographic resolution. In fact, the majority of the achievement in resolution improvement has been accomplished by decreasing the lithography ”difficulty” factor, k_1. For modern lithography systems, k_1 is pushed well below Rayleigh’s 0.61 through the use of high-contrast photoresist and a host of resolution enhancement techniques.

As features grow smaller in proportion to the wavelength, diffraction effects become more noticeable. Features with sharp corners on the mask are printed with rounded corners. Extremely narrow features will not print if they are isolated (since they do not benefit from diffracted intensity spillover from neighboring lines as they would if the lines were densely packed). To combat these effects, mask engineers employ optical proximity corrections (OPC). Sharp corners can be achieved by adding ”serif” or ”hammerhead” structures (**Fig. 2.3**) to the features on the mask, and the iso-dense bias can be similarly relieved by adding scattering bars (sub-printable features) near isolated features. When the mask pattern is printed, these additional features serve to create a sharper pattern.

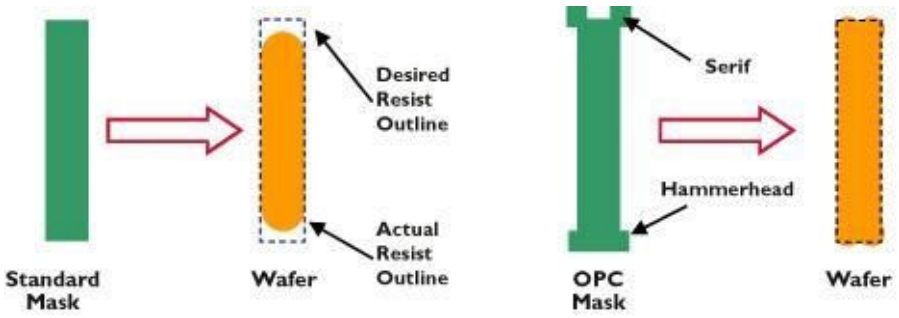

Figure 2.3. Standard optical proximity corrections. On the left, no OPC adjustments have been made. On the right, the serif and hammerhead features enhance the printing of the desired feature [2.2].

Other resolution enhancement techniques include phase-shifting masks, off-axis illumination, and multiple exposures. Virtually every component of the exposure system has been explored for sources of potential reduction in k_1 (**Fig. 2.4**). Not surprisingly, there is a price to pay for employing these techniques to enable the printing of sub-wavelength features: increased variability in the lithography process.

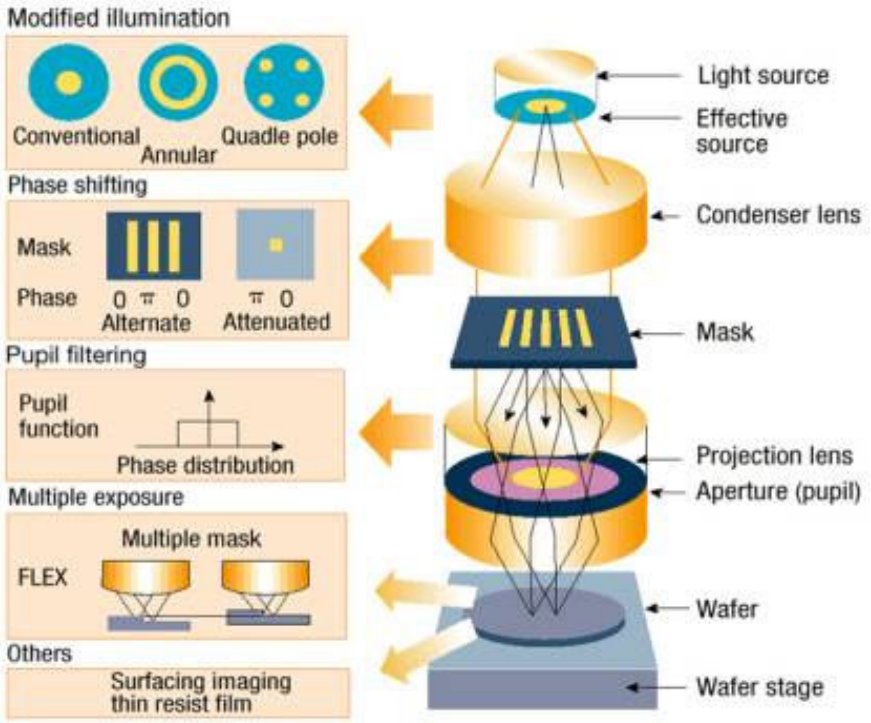

Figure 2.4. A standard projection lithography system, with examples of resolution enhancement technologies [2.3].

2.2.1 Resist, Bake, and Development

Photoresist has three main components: a resin (base material), a photoactive compound (PAC), and a solvent that controls the structural properties of the substance, particularly viscosity. In positive resists, the PAC behaves as an inhibitor, nominally preventing the resist from dissolving in developer solution. Exposure to light triggers a chemical reaction which changes the PAC into a sensitizer—a compound which enhances the dissolution rate in developer by breaking the resin structure material down. Resist contrast is determined by the sensitivity of the resist to changes in exposure dose. The narrower the range of exposure energies between zero resist removal and total resist removal, the higher the resist contrast. In turn, the higher the resist contrast, the sharper the line edge in the printed feature. Resist resolution—the minimum feature size which may be printed in the resist—also depends on resist contrast, since the contrast of the aerial image (i.e., the light before it interacts with the resist) must exceed the contrast of the resist in order for the pattern to print [2.4].

Photoresist is applied to the wafer with a spin-coating process, which yields an exceptionally uniform film across the wafer. However, from wafer to wafer and lot to lot, variability in spin speed, resist viscosity, and adhesive properties between the resist and substrate lead to film thickness variation. This variation will in turn lead to variation in the printed CD—the same amount of exposed intensity will yield wider features for a thicker resist layer.

Below about 250 nm, traditional resists begin to strongly absorb incident light, leading to non-uniform exposures through the depth of the film. At the DUV (248 nm)

lithography technology node, this absorptive property rendered traditional resists virtually obsolete, and new chemically amplified resists (CAR) were developed. Chemical amplification is achieved by an agent contained in the resist which greatly increases the potency of the photochemical process. The photoacid generator (PAG) of a CAR is designed so that an incident photon will create a cascade of chemical reactions that catalyze the scission of the resin [2.5]. The resist sensitivity is dramatically increased, but relies heavily on thermal activation, which is carried out in a post-exposure bake (PEB) step. The resulting latent image (prior to development) in the resist depends strongly on the thermal dose—both PEB temperature and duration are key factors. Therefore, PEB bake stations must be designed to place the wafer perfectly level on the heated surface so that the same thermal dose is received at all points on the wafer; also, the nominal temperature must be carefully maintained, particularly difficult since the bake plate encounters a cold, room-temperature wafer at the beginning of each bake cycle. At a nominal temperature of roughly 100^{o}C, deviations of a single degree will lead to ~5 nm deviations in resulting feature width.

There are two other bake steps in a standard lithography flow: the post-application bake (PAB) and an optional hard bake. The PAB step serves to drive off some of the solvent in the film following spin-coat, serving to create a firmer film prior to exposure. The hard bake step follows development, and is intended to drive off any remaining solvent in the remaining resist to strengthen the masking features prior to the etch process. Since neither step has a strong effect on the resulting pattern, they generally do not receive much attention from a process control standpoint.

Although the resist feature width is strongly dependent on the development process at the beginning of the develop step, this dependence quickly reduces to well below the sensitivity level of exposure, focus, resist thickness, and PEB characteristics. Therefore, lithography modules are designed such that the nominal development time is much longer than the time period over which the feature is sensitive to this step, ensuring that slight variations will not contribute much variability in the process. Thus, the development step is a secondary concern as a source of variability, and is commonly ignored in the process control setting.

2.3 Etch

Once the desired pattern has been formed in the photoresist layer, the etch process is used to transfer the photoresist pattern into the layer of underlying material. Etch processes fall into two broad categories—wet etching and dry etching. Wet etch processes, where the entire wafer is immersed in a solution that reacts with and thereby removes the exposed parts of the underlying material, tend to be used only for non critical processes due to the inherent difficulty in controlling the properties of the etch. For example, wet etches are usually isotropic—the etch rate is uniform in all directions, so features undercut the regions defined by the photoresist mask. In addition, wet etch processes introduce a high risk of defects (via particulate contamination), making these processes unsuitable for defining small features. Dry etching processes, on the other hand, are relatively clean and can be tuned to generate the desired selectivity, etch rate, and anisotropy. The dry etch is performed using a plasma in a controlled, vacuum environment called a plasma chamber (shown in **Fig. 2.5**). The plasma is created using

an electric field to break down an inert gas into electrons and ions; the ions are accelerated toward and bombard the cathode to generate secondary electrons, which are in turn accelerated back toward the anode and thus collide with the gas to create more ions. The science and engineering of maintaining a controlled plasma glow discharge is complex and beyond the scope of our needs in this introduction [2.4].

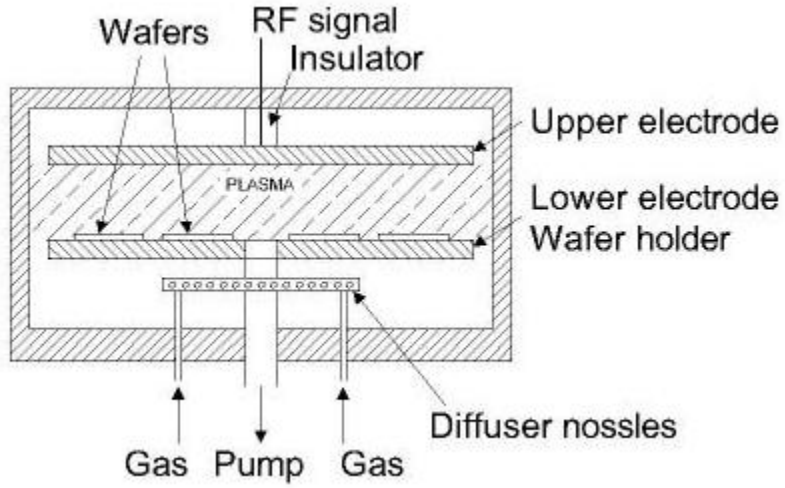

Figure 2.5. Illustration of a multi-wafer reactive ion etch plasma chamber [2.6]. An electric field is set up between the upper and lower electrodes to create the plasma; as feeder gas is diffused into the system via the designated nozzles, the plasma breaks the gas down into ions, electrons, neutrals, and radicals. Reactive species diffuse to the surface of the wafers mounted on the lower electrode, and with the aid of ion bombardment, react with and remove the exposed layer. Byproducts are desorbed from the wafer surface and pumped out of the system.

To begin the etching process, a feeder gas is then introduced into the chamber and is broken down by the plasma into ions, radicals, electrons, and neutrals. Etching then occurs via two mechanisms: first, the chemical reactant species diffuse to the surface of the wafer and react with the exposed material. Byproducts may then be desorbed, diffused away from the wafer, and transported out of the plasma chamber. Second, physical bombardment by ions (driven into the surface of the wafer by the electric field

used to set up the plasma) causes damage to the surface of the exposed film, both promoting the chemical reaction and the removal of byproducts. In some plasma etch processes (referred to as ion milling processes), material removal almost completely depends on the mechanical energy of ion bombardment—but for front-end pattern transfer process, a balance between chemical and mechanical etch is carefully struck.

Since the plasma may be easily started and stopped, it is much easier to start and stop the dry etching process than a simple wet etch process. In addition, the vertically-oriented bombardment of ion species (arising from engineering the plasma environment to have a mean free path large in comparison to the dimensions of electrode separation) gives rise to high anisotropy, or nearly vertical sidewalls in the defined features. Finally, since there are many fewer particles existing in the plasma environment than there are in an immersion environment, the defect problem is strongly reduced. However, since the etch rate will depend on local concentrations of the reactant species as well as the removal rate of the reaction byproducts, IC layout dependent microloading effects can arise during the etch process. These effects and other sources of variation will be addresses in the next chapter.

References

[2.1] M. Switkes and M. Rothschild, "Resolution enhancement of 157-nm lithography by liquid immersion," in Optical Microlithography XV, A. Yen, ed., Proc. SPIE 4691, pp. 459–465, 2002.

[2.2] "Asml masktools division." http://www.masktools.com/content/mttechno.htm.

[2.3] T. Ito and S. Okazaki, "Pushing the limits of lithography," Nature 406, pp. 1027–1031, 2000.

[2.4] S. A. Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, New York, NY, 2001.

[2.5] M. Madou, *Fundamentals of Microfabrication*, CRC Press, Boca Raton, FL, 1997.

[2.6] *MEMS and Nanotechnology Clearinghouse*, http://www.memsnet.org/mems/.

Chapter 3

Complete Characterization of Spatial Variation in Critical Dimensions

This chapter covers the bulk of experimental work carried out with the goal of generating a complete characterization of spatial variation of critical dimensions. First, the principal causes of CD variation as well as associated process control opportunities are reviewed. Next, a historical gate length variation study aimed at exploring long range ($1mm$—$2cm$) variation is described. Finally, novel electrical linewidth metrology test structures for measuring short-range ($0.2\mu m$—$1mm$) spatial variation in gate length are presented, including a description of design, characterization, and results.

3.1 Sources of CD Variation

Manufacturing-induced variation in critical dimensions has rich structure. Using a wide range of metrology tools, this variation can be thoroughly characterized and decomposed into several deterministic components, including wafer-to-wafer variation, across-wafer variation, within-die variation, and pattern-dependent variation. Finally, spatial correlation analysis can be performed on the remaining, residual distribution to

determine whether it is truly random, or whether it still contains some systematic spatial information.

3.1.1 Wafer-to-Wafer Variation

Wafer-to-wafer variation arises from variation in the state of the manufacturing tool. Over time, the conditions of a given process may drift so that wafers passing through a given process step near the beginning of a lot undergo a slightly different process than those processed near the end. In the photolithography module, temporal drift in conditions of "best" focus and dose requires constant adjustment. This is a problematic phenomenon since drift in the width of a printed feature can be due to drift in either focus or dose, and as such a substantial research effort is put into determining the optimal process settings as they change over time. Competing techniques include searching for particularly-sensitive patterns within a given design that may allow the engineer to separate dose and focus-induced variation [3.1] and the design of external test structures that sit in the scribe lanes and provide accurate measure of focus and dose at the edge of the field [3.2]. Similar sources of wafer-to-wafer variation exist in the etch and CMP modules; however, due to major advances in general fab quality control, principally through automated statistical process control, wafer-to-wafer variation can be considered a lesser source of concern than across-wafer and within-field variation.

3.1.2 Across-Wafer Variation

Across-wafer variation has multiple sources. Every step of the patterning process during which the wafer is processed in whole will contribute to the across-wafer component of variation. Within the lithography sequence, the post-exposure bake (PEB)

step can prove to be the greatest variation culprit if it is improperly characterized and monitored. Since chemically amplified resists are strongly affected by PEB, resist temperature sensitivities on the order or 7-10nm/°C [3.3] must be tolerated. This sensitivity exists by design—the PEB step causes thermally catalyzed deprotection of the exposed areas of the resist at a rate that is strongly dependent on bake temperature and time. As an upshot of this strong sensitivity, any thermal gradient in the PEB bake plate is transferred into the thermal profile in the resist layer, and will subsequently lead to a strong deterministic across-wafer component of variation. The effect is demonstrated in **Fig. 3.1**, which shows the PEB bake plate temperature profile and the resulting across-wafer CD distribution for two wafers processed on two separate, poorly-controlled bake plates. In areas where the bake plate is relatively cool, the critical dimension is larger than average, and in areas where the bake plate is relatively hot, the CD is smaller than average. Fortunately, this PEB temperature gradient effect is predictable and stable over long periods of time (several weeks). In addition, the past five years have seen the development and widespread use of wafer-embedded temperature sensors to exhaustively characterize bake plates that incorporate multiple heating zones. By monitoring the bake plate thermal profiles and performing adjustments to individual zones, bake plates have been made more uniform both internally and across the range of plates that accompany each tool [3.5]. However, control of the transient heating sector of the bake plate's thermal trajectory still presents a significant impact on CD variation, and a similarly significant control problem [3.3, 3.6].

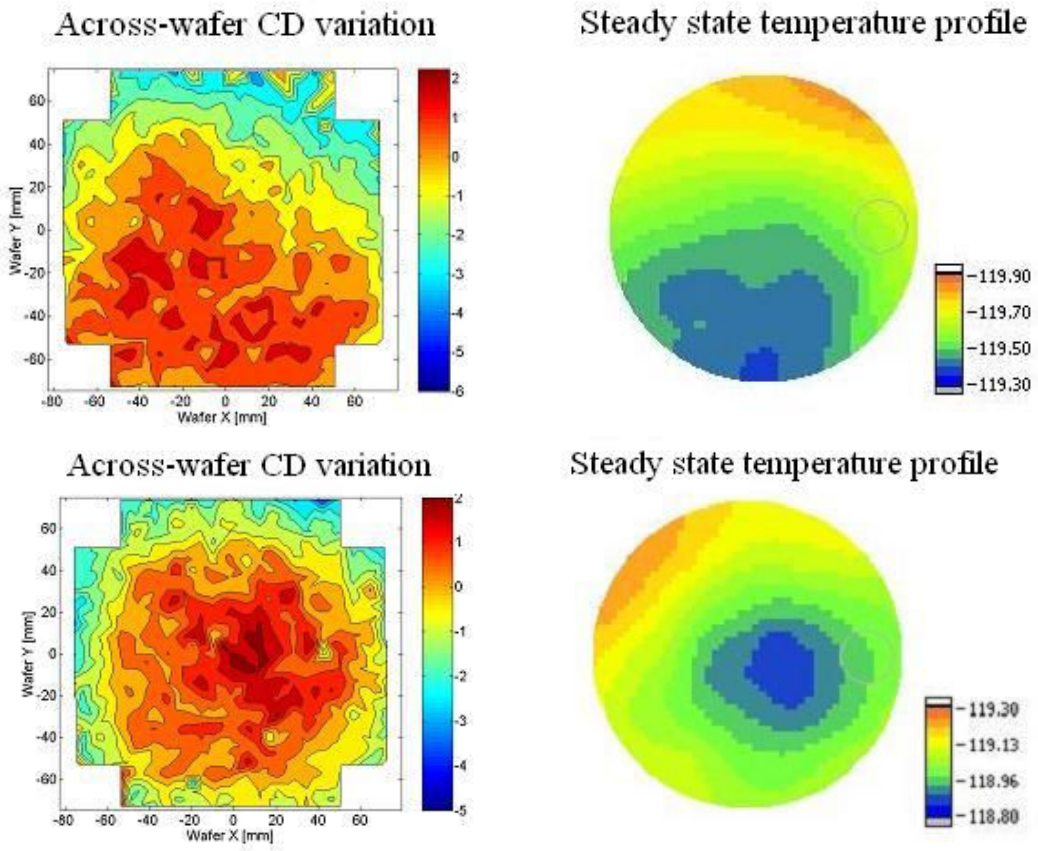

Figure 3.1. Across-wafer CD variation diagrams and corresponding steady-state thermal fingerprints for two single-zone bake plates [3.4]. The left-hand plots the CD deviation from wafer-average CD, which the right-hand plots capture the temperature profile of the bake plate during the stead state region of the bake trajectory. A systematic relationship between the two is clearly evident.

Additional sources of across-wafer in the lithography process exist, but they are relatively small in comparison with PEB variation. The spin-application of resist and anti-reflective coatings as well as the developing of the resist may contribute a slight radial variation, but they are negligible in comparison with the PEB effect.

Finally, the plasma etching step that serves to define features in the underlying layer contributes a substantial dome-shaped variation due to imperfect uniformity in the etching tool. This source of variation is comparable in size to the PEB variation

component. Since plasma uniformity is much more difficult to control, efforts have been made to characterize the etch variation and then feed that information back to the PEB step, where pre-compensating variation may be intentionally introduced to offset the etch step [3.7].

3.1.3 Within-Die Variation

Two major sources of within-die variation exist: mask errors and variations caused by the lithographic scanner. Mask errors are introduced during the writing of the reticle, in both the nanometer range (from write noise, such as resist charging, which is caused by operating the write-tool at the edge of its capabilities) as well as the 10nm-to-100nm range (from etching non-uniformity and write-tool stitching) [3.8]. These mask errors are transferred onto each lithographic field equally during the printing of device wafers. Although mask errors cannot be corrected once the reticle has been manufactured, they can be characterized through rigorous inspection.

Within-die variation caused by the scanner tool itself is caused by two separate effects related to the operation of the tool. Since a relatively thin exposure slit is scanned with a certain dose across the entire field, the variation can be separated into along-slit variation and along-scan variation. The along-slit variation arises from imperfections in the lens and illumination system, and is usually symmetric in nature due to the symmetry of the components in the exposure tool. The along-scan variation is much smaller in magnitude than the along-slit component, and arises from a combination of lens imperfections and systematic variability in the delivered dose along the scan. In

combination, the two sources of exposure tool-caused within-die variation leave a fingerprint that is relatively consistent across a range of scanner tools (**Fig. 3.2**).

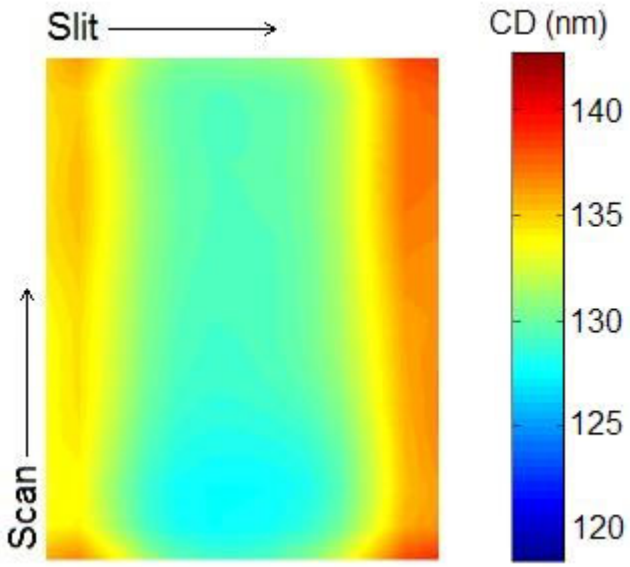

Figure 3.2. Within-die CD variation of the lithographic scanner. The characteristic variation fingerprint along the slit (horizontal) and along the scan (vertical) are illustrated [3.9].

Substantial progress has been made in controlling within-die variation caused by the exposure tool; the along-slit and along-scan components are addressed separately [3.9]. For along-slit variation, a graduated filter has been proposed to offset the relatively consistent variation in the lens and illumination systems for each tool. To combat along-scan variation, the energy of the laser pulses used to generate the exposure light is tailored to offset the naturally occurring along-scan dose variability. Combined, these corrections provide significant detuning of the original total variation, reducing the 3-σ level of variation by 35%, or 6.5nm in this case. However, this type of exhaustive within-die process control is costly since each tool must be characterized independently and would require periodic retuning to optimize the correction.

3.1.4 Pattern-Dependent Variation

As device sizes shrink, semiconductor processes are pushed to the extremes of their capabilities, and actual device geometries subsequently become subject to pattern-dependent effects. In the lithography module, off-axis illumination and variability of aberration effects give rise to "forbidden" pitches—duty cycles of patterning that are especially vulnerable to variation [3.10]. Also, lithography proximity effects (where "spill-over" light from neighboring features contributes to the exposure of another feature) give rise to a bias between isolated and dense features as well as variation in the simple geometry of a single feature (e.g., line-end shortening). To combat these effects, lithographers devote tremendous effort to optimizing illumination settings, exposure parameters, and reticle enhancement techniques (RET) such as optical proximity corrections (OPC) and sub-resolution assist features to restore the printed geometry to its intended design. Designs are also tailored to present a limited range of desired lithography features to be printed: "pitch selection" and rigid orientation of features, also known as "Restrictive Design Rules," may be deemed necessary in the not too distant future [3.8].

Pattern-dependent effects are not just limited to the lithography module. In the etch module, density-dependent effects are seen, both based on total exposed area ("loading effect") and local variations in the pattern density ("microloading"); reactant species are depleted at different rates in isolated areas than they are in dense areas, influencing the local etch rate [3.11, 3.18]. In the chemical-mechanical polish module, dishing and erosion effects are largely dependent on local geometries [3.12]. Even in the implantation module, variability in the overlaying polysilicon geometries can lead to

variation in dopant diffusion, since areas with high density of polysilicon heat to higher temperatures during a rapid thermal anneal [3.13]. In all these cases, pattern regularity is enhanced through use of dummy fill and/or reticle biasing to combat pattern-dependent variation.

3.1.5 Spatial Correlation

According to Pelgrom's model [3.14], the variance in mismatch for a given parameter P of two rectangular devices varies depending on the size of the devices and their separation distance as:

$$\sigma^2(\Delta P) = \frac{A_p^2}{WL} + S_p^2 D_x^2 \tag{3.1}$$

The first term in the model accounts for purely random (or "white noise") variation that scales with the inverse of the area of the device, while the second term accounts for variance in parameter mismatch (systematic) that scales with the square of the separation distance. Pelgrom's underlying assumption is that the distribution of parameter P is a smooth quadratic function of distance; the random selection, then, of two sampling points at a fixed separation distance D_x along this function gives rise to the position-dependent term in **Eq. 3.1**. This model implicitly suggests that spatial correlation is an inherent property of semiconductor devices, since devices situated closely together will have less variance in parameter mismatch than devices separated by relatively long distances. However, any rigorous discussion of spatial correlation would require that the underlying distribution of the parameter P be stationary, and in Pelgrom's work, the distribution of P was assumed to non-stationary. As such, a mathematically rigorous extension to true

spatial correlation is not easily gained from his work. Our goal is a more general, comprehensive formulation of spatial variability that includes true spatial correlation as one component; therefore, we will perform a novel examination of the role of "true" spatial correlation in this work, better suited to measuring variability in delay of a critical path, rather than mismatch in a single device parameter.

Following our new objective, it is unclear whether pure spatial correlation is a significant property of parameter variation. However, it is certainly worth investigating because the implications of spatial correlation are significant. Consider the variance of the sum of n identically, normally distributed, random samples with a fixed correlation ρ among all pairs:

$$\sigma^2_{tot} = n\sigma_{ind}^2 + 2\rho\left({}_nC_2\right)\sigma_{ind}^2 = \left[n + \rho(n)(n-1)\right]\sigma_{ind}^2 . \tag{3.2}$$

Using this relationship, we see that when $\rho = 1, \sigma_{tot} = n\sigma_{ind}$, whereas when $\rho = 0$ $\sigma_{tot} = \sqrt{n}\sigma_{ind}$. The difference between the two conditions is clearly substantial, particularly when n is large. This formulation compares nicely with the calculation of the delay of a critical path where the parameter values of each gate vary randomly with some correlation. For example, if we have a chain of n identical inverters with a distribution of delays with equal mean, variance σ, and some correlation ρ, then the variance of the total delay of the chain will be very close to the result given in **Eq. 3.2**. To illustrate, consider the example of the delay of a 4-stage inverter chain (**Fig. 3.3**) with gate lengths randomly drawn from a normal distribution subject to some level of correlation. If we simulate several hundred cases, in which we perform a randomly drawn gate length assignment to each gate in the chain for each case, we can determine how the variation in delay depends on the correlation.

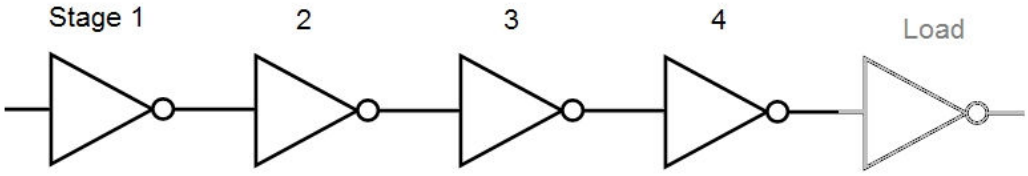

Figure 3.3. Inverter chain canonical circuit used for exploration of general effects of correlation. Gate length values are assigned randomly to each stage (identically to both NMOS and PMOS transistors within each stage, scaled to the standard inverter sizing bias) such that the correlation between any pair of stages is ρ. This method is a standard treatment of correlation in related literature; we will see that this method needs to be expanded to allow for variable values of ρ for different separation distances.

The result can be compared graphically with the result of **Eq. 3.2**, as shown in **Fig. 3.4**. This plot shows that the results measured in SPICE do not strictly match those predicted by theory; this discrepancy is due to a violation of the assumption of independence in the inverter chain submitted to SPICE. Since the variable size of the load on a given gate will also impact the delay of that gate, there is indeed some interaction between the delays of consecutive gates even if their critical dimensions are actually independent. Therefore, the reduction in delay variation predicted for completely independent stages is slightly overestimated in comparison with the physical result.

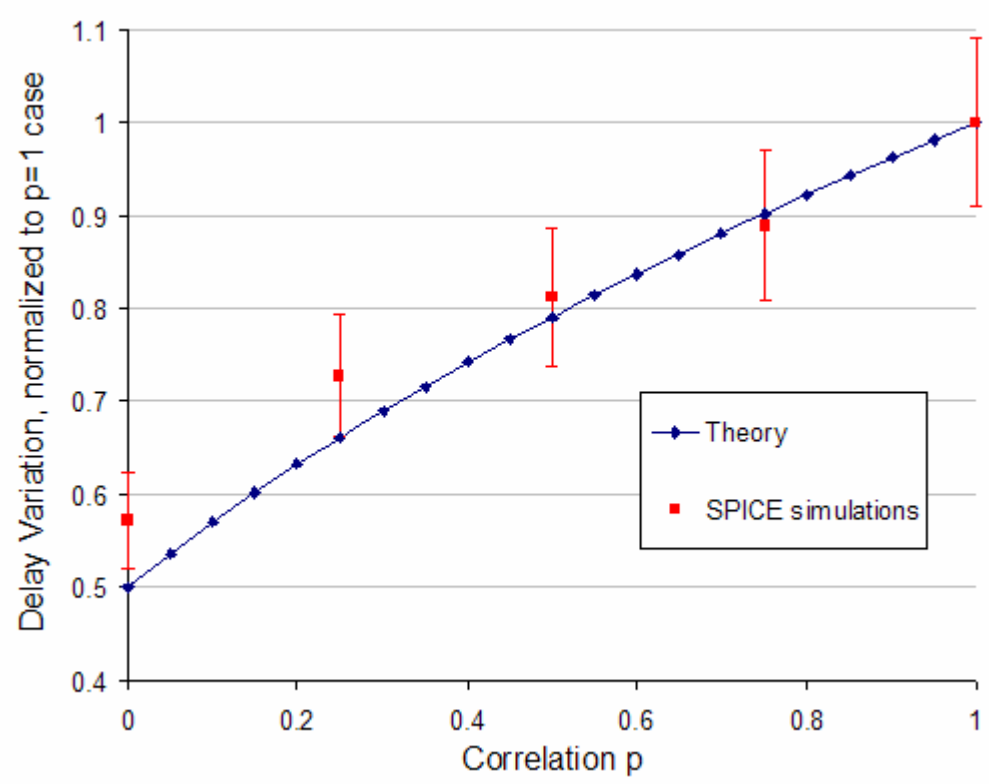

Figure 3.4. Comparison of delay variation of a 4-stage inverter chain with randomly drawn gate lengths subject to correlation ρ as measured via SPICE simulation with theoretical result if stages are completely independent. It is evident that the reduction in variation as correlation is reduced given by statistical theory overestimates the actual realized reduction in the delay variation of the inverter chain. This discrepancy is attributed to the dependency of a given gate's delay on the size (randomly drawn) of the following gate. The error bars on the SPICE simulation data points indicate 95% confidence intervals around the value estimated through 200 Monte-Carlo simulations.

Previous studies that have examined spatial correlation usually end at this basic level of analysis; they compare the result under perfect correlation with the result under zero correlation. However, in a rigorous analysis of spatially distributed samples, the spatial correlation will not be fixed at one value, let alone fixed at the extremes of 1 or 0. For a more rigorous treatment, we need to perform calculations involving the covariance matrix, allowing for variable values of ρ for different separation distances. Using this more general approach, the variance matrix (a column array, where the i^{th} entry

corresponds to the variance associated with the i^{th} sample) of the n equidistant samples arranged along a line, is given by the relation:

$$variance\ \ matrix = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \sigma_3^2 \\ \vdots \\ \sigma_n^2 \end{bmatrix} \tag{3.3}$$

Here, ρ_1 is the correlation between samples separated by one length unit, ρ_2 is the correlation between samples separated by two units, and so on. To calculate the total variance of the sum of the n samples, assuming all the individual variances are the same, we simply add the components of the resulting column vector to get:

$$\sigma_{tot}^2 = \left(n + (n-1)(2)\rho_1 + (n-2)(2)\rho_2 + \ldots + 2\rho_{n-1}\right)\sigma_{ind}^2 \tag{3.4}$$

From this analysis, we see that the correlation between samples situated close together is weighted more strongly in the total variance expression, which is intuitive since there are more pairs of samples separated by those closer distances. Because of this significant dependency on the precise structure of spatial correlation, it is evident that a thorough analysis of spatial correlation would be useful.

3.2 Long-Range (1*mm*—2*cm*) Variation

To begin a full characterization of CD variation, the first step taken was to revisit an experiment previously carried out [3.15] by Jason Cain, in which a quantitative analysis of across-wafer and within-die variation was performed through the collection of CD

measurements separated by relatively long-range (1mm—2cm) distances. The previous work focused on optimizing sampling plans to adequately capture these deterministic components of variation; in this work, however, closer attention is paid to the residual distribution of variation once the across-wafer and within-die components have been removed.

3.2.1 Historical Gate Length Study

In the previous project, exhaustive critical dimension (CD) measurements were performed using electrical linewidth metrology (ELM) on a full 200mm wafer processed through an industrial 130nm manufacturing process [3.15]. The linewidth of a poly-silicon "gate" is measured using a Kelvin structure, by passing a precisely calibrated current through the gate and measuring the voltage across a subsection of the gate as shown in **Fig. 3.5**.

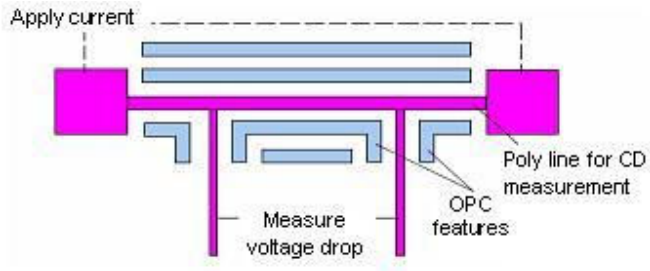

Figure 3.5. Kelvin test structure for ELM measurement. Current is pushed between the outer pads, and the voltage drop across a known length of line composed of the material of interest. This measured voltage drop, combined with sheet resistance measurements from neighboring Van der Pauw structures, can be used to extract linewidth. [3.15]

The electrical measurement yields a numerical value with a sizable (~40nm) negative offset from values measured using more standard metrologies such as CDSEM or

scatterometry, but the ELM measurements have been shown to have greater precision than other metrologies and exceptional speed. Due to this speed of measurement, large quantities of data may be taken—for example, this analysis makes use of 280 measurements per die over 35 dice, a total of 9800 measurements (**Fig. 3.6**).

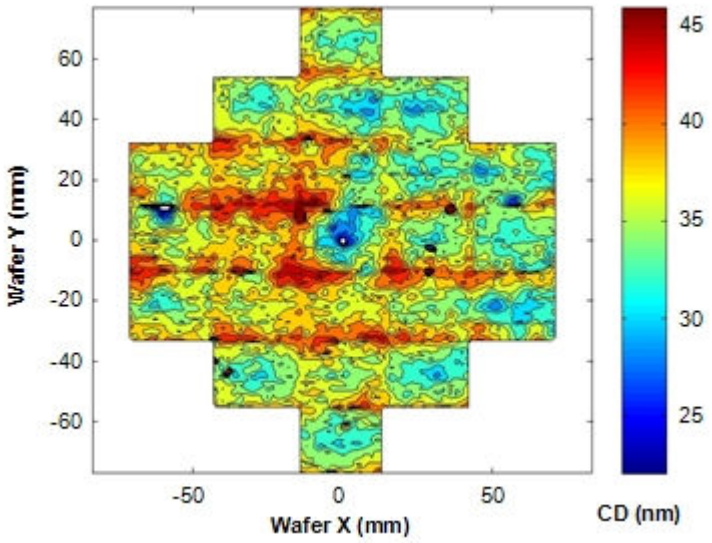

Figure 3.6. Full-wafer CD measurements [3.15]. Each field contains 280 measurements, for a total of 9800 measurements per wafer. The measured CDs, taking an average value of roughly 35nm, reflect the expected large, negative offset (~40nm) of the ELM measurement methodology.

3.2.2 Decomposition of Variance

We would now like to determine how various forms of process control in the lithography cell might impact the spatial correlation dependence. To accomplish this goal, we begin by decomposing the CD measurements into five components: the average CD (μ), mask errors (*mask*), within-field systematic variation (*WIF*), across-wafer systematic variation (AW), and random variation (ε):

$$CD_i = \mu + mask_i + WIF_i + AW_i + \varepsilon_i \tag{3.5}$$

where the mask errors are measured directly from the reticle used to generate the patterns, and the *WIF* and *AW* components are extracted from the data by least-square fitting to a 2nd-order polynomial

$$fit = \beta_1 x^2 + \beta_2 y^2 + \beta_3 x + \beta_4 y \tag{3.6}$$

Starting with the full-wafer dataset, the mask errors are first removed. Then, the average within-die CD variation fingerprint that is caused by non-uniformity in the scanner tool may be extracted and modeled (**Fig. 3.7**). It is clear from this analysis that there exists a strong systematic within-field variation component. When this within-field variation component is removed from the full-wafer CD measurements, there is an obvious across-wafer systematic variation component, which can also be modeled well with a second-degree polynomial with coefficients estimated through least squares regression (**Fig. 3.8**).

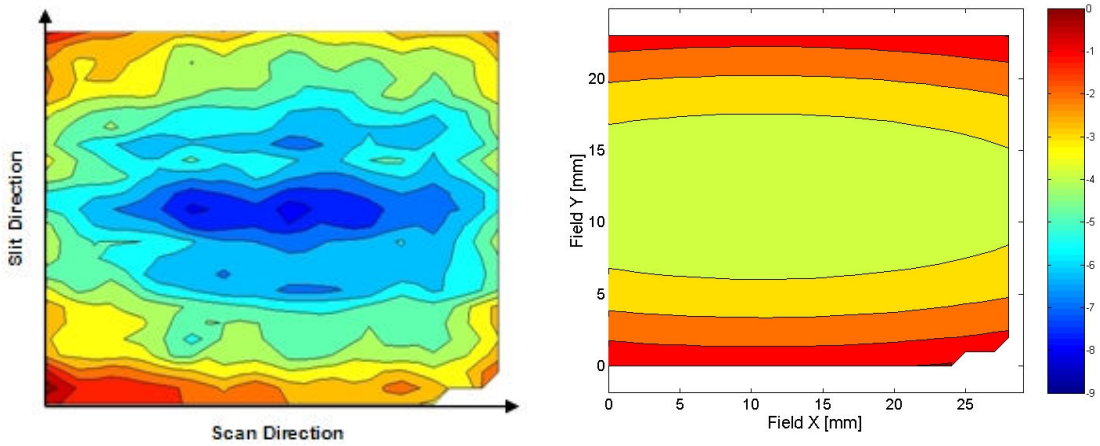

Figure 3.7. Average within-die CD fingerprint (left) and polynomial model of within-field variation (right). Cross-sections of this model dictate the nature of horizontal and vertical spatial correlation [3.15].

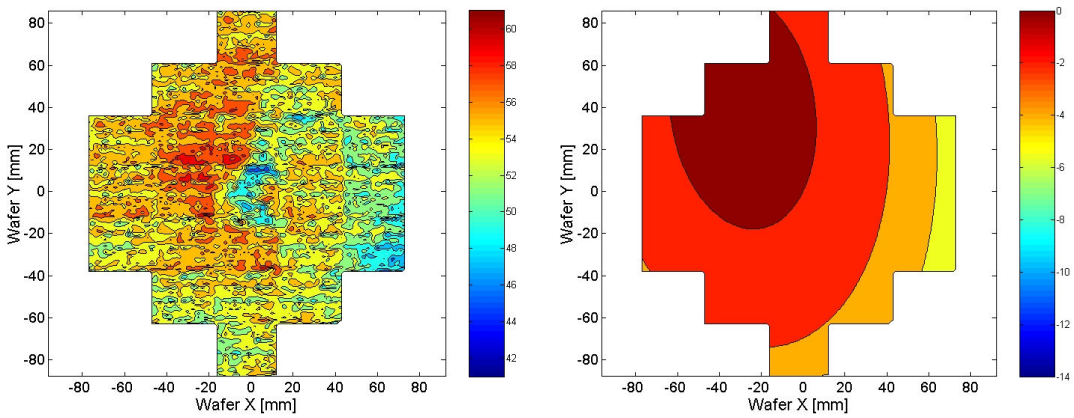

Figure 3.8. Full wafer CD map with average within-die CD fingerprint removed (left) and polynomial model of across-wafer variation (right) [3.15].

Once the *WIF* and *AW* components of variation that were modeled with second-degree polynomials have been removed from the initial full-wafer dataset, the new, residual full-wafer dataset can be examined again to search for remaining spatial dependency. More specifically, one would look for any remaining systematic spatial variability, and, as shown in **Fig. 3.9**, the residual full-wafer data set still contains some clear systematic variation. If the dataset is indeed non-stationary, then the statistical assumptions required to calculate spatial correlation will be violated.

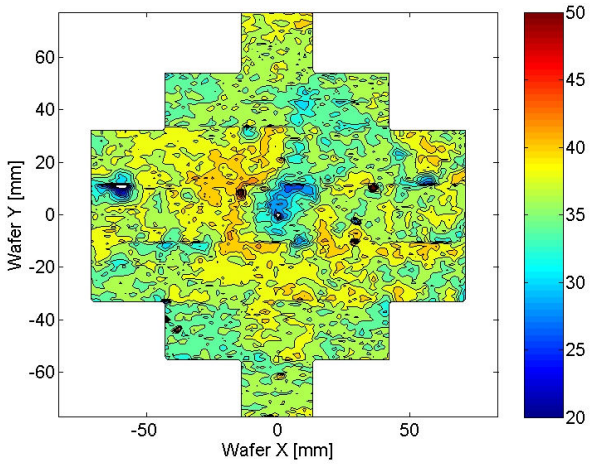

Figure 3.9. Full-wafer CD map with *WIF* and *AW* polynomial model systematic variation components removed.

It appears that the average CD for each field may still vary from field to field. Under suspicion that this variation is caused by errors in dose control from shot to shot during exposure, we can apply a field-by-field correction term [3.9] that would simulate the reconciliation of this source of variation (**Fig. 3.10**).

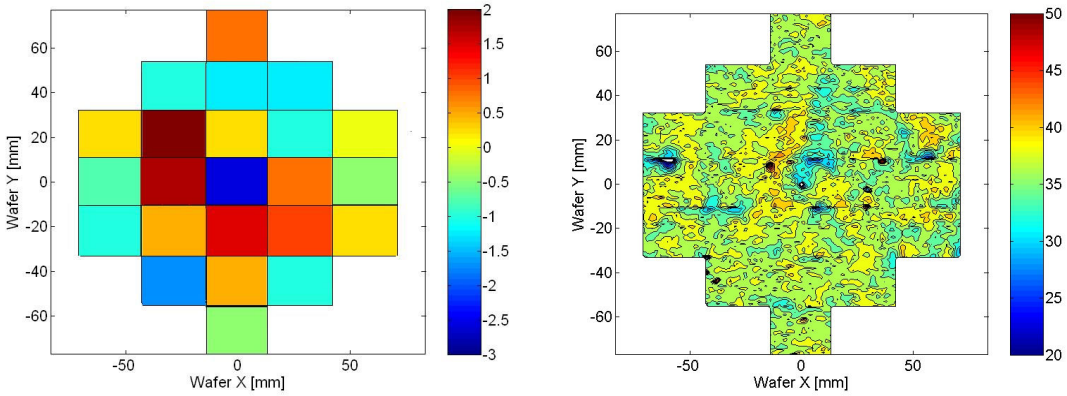

Figure 3.10. Die-to-die dose-error correction offsets (left) and resulting full-wafer CD map when these corrections are applied to the data set (right).

At this point in decomposition of variance, every source of systematic variation that we could hope to control through adjustments to the process has been identified. The residual distribution is roughly stationary—at least, as stationary as we can render it without over fitting the data for systematic dependencies. Therefore, we will next turn our attention to the analysis of spatial correlation.

First, each CD measurement from the residual distribution is standardized against the entire residual CD distribution using the relation

$$z_i = \left(x_i - \hat{\mu}\right)/\hat{\sigma}\,, \tag{3.7}$$

where x_i is the i^{th} residual CD and $\hat{\mu}$ and $\hat{\sigma}$ are the average and standard deviation of all residual CD's. Next, the spatial correlation approximation can be calculated and plotted as a function of separation distance using

$$\rho_{jk} = \left(\sum z_j * z_k\right) / n \,. \tag{3.8}$$

That is, for each separation distance *l* afforded by the density of the measurements, all pairs of points (*j* ,*k*) separated by *l* that fall within the same die are included in the summation expressed in Eq. 3.8.

By sampling across all separation distances afforded by the dataset, an autocorrelation diagram can be created, displaying the spatial correlation measured between all pairs of measurements at a certain separation distance for each separation distance. The resulting spatial autocorrelation functions for both the original CD dataset as well as the residual CD data set are plotted in **Fig. 3.11** (displaying correlation versus separation distance in both the horizontal and vertical directions of separation distance). It is clear that the shape of the within-field systematic spatial variation component heavily influences the shapes of the autocorrelation functions calculated for the original dataset (upper curves, labeled "before spatial modeling"). The obviously counterintuitive increase in spatial correlation for large separation distances indicated in that curve illustrate the consequences of performing correlation analysis on a non-stationary dataset. On the other hand, once the deterministic variation components have been removed, the autocorrelation functions calculated from the residual dataset show that very little true spatial correlation exists for this distance scale (bottom curves, labeled "after spatial modeling"). Indeed, one might assume that even the little remaining spatial correlation might simply be an artifact of imperfect removal of the above-mentioned deterministic variation. That is, in the die-scale regime, spatial correlation is virtually entirely an artifact of systematic variation.

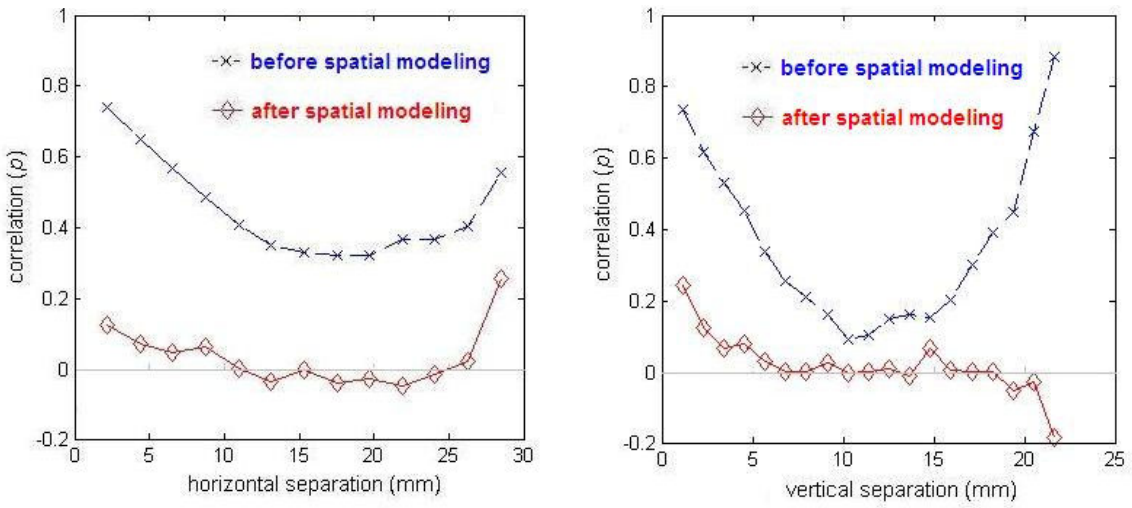

Figure 3.11. Spatial correlation dependences with full across-field, across-wafer, and field-by-field spatial modeling applied for horizontal (left) and vertical (right) separation.

One feature of the autocorrelation functions of the residual CD distribution is the up-tick in spatial correlation for the shortest separation distances afforded by our data set. This up-tick suggests that while there is little spatial correlation in the die-scale separation distance range, there may be substantial spatial correlation in the short-range, 0.2µm to 2mm range, which is likely the crucial path length range to consider for common critical paths. (The apparent correlation swing at large correlation distances is due to estimation noise, since large distance pairs are not as numerous in our population sample.) Unfortunately, since the minimum separation distance between test structures in this historical study is roughly 1mm, a separation distance range of great interest is unavailable. To fill in this gap in understanding, new test patterns were designed to target the short-range ($0.2\mu m$—$1mm$) regime of CD variation.

3.3 Short-Range (0.2*μm*—1*mm*) Variation

3.3.1 Serpentine ELM Test Structures

In order to capture a detailed picture of spatial variation spanning distances in the micron-scale, dense measurements are needed. To collect such dense measurements, a novel serpentine ELM test structure was created and deployed to a 90nm foundry service for manufacture. The base-case ELM test structure (shown in **Fig. 3.12)** combines a series of polysilicon lines (measurable by ELM method) packed at maximum density.

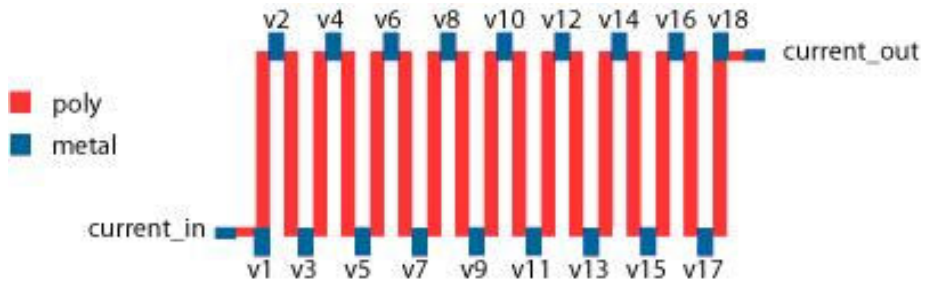

Figure 3.12. Base-case micron-scale ELM structure (not to scale). The serpentine structure, composed of polysilicon, can be considered as 17 separate Kelvin test structures. Current can be forced through each length of polysilicon independently, and the appropriate voltage drop may be measured using the denoted voltage taps. By combining the measured resistance with sheet resistance data, the linewidth of each segment may be independently captured.

Each polysilicon line is designed to have the minimum available linewidth for the process (90nm design rule, 100nm as drawn in layout), and has a length of 22μm to reconcile any local variation in width due to line-edge roughness. The structure is also designed at minimum pitch (280nm) allowed by process design rules. By applying a current through the entire serpentine structure and measuring the voltage between

successive voltage measurement nodes (v1—v2, v2—v3, and so on) and dividing by the value of the applied current, the resistance of each line (R_p) may be extracted.

Next, the local sheet resistance is measured from the Van der Pauw structure that accompanies each test structure. The Van der Pauw structures, located adjacent to each test structure, were designed in Greek cross style in order to minimize finite contact errors [3.16] as shown in **Fig. 3.13**. The sheet resistance is given by the relation:

$$R_s = \pi(R_{34,12} + R_{13,24})/(2\ln 2) \tag{3.9}$$

where

$$R_{34,12} = (V_3 - V_4)/I_{12} \quad \text{and} \quad R_{13,24} = (V_1 - V_3)/I_{24}, \tag{3.10}$$

with the subscripts corresponding to the contacts to the structure as labeled in **Fig. 12**.

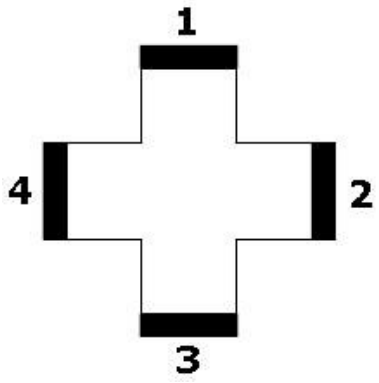

Figure 3.13. Greek Cross Van der Pauw structure [3.16]. By passing current between two contacts of the structure and measuring the voltage drop between the remaining two contacts, the sheet resistance of the material comprising the structure may be calculated, using Eq 3.9 and 3.10.

Finally, to extract the linewidth of the polysilicon line, we make one final geometric calculation:

$$linewidth(nm) = R_s * 22000nm / R_p \tag{3.11}$$

The base-case serpentine ELM structure has a total range (separation distance as measured between the outermost measurable lines) of 4.76μm. To explore the entire

range that was missing in the historical long-range CD variation study, though, it is necessary to sample up to a separation distance of 1mm. Therefore, to extend the spatial frequency range of the test structure, variants of the base-case ELM structure were designed to include the addition of "dummy" lines that increase the separation between the measurable poly-silicon lines without changing the overall geometric arrangement (minimum pitch, minimum linewidth). In this way, the results from all the structures could be combined into a single dataset since the lithographic pattern density is consistent across all structures, whether they contain dummy lines or not. **Fig. 3.14** shows the first dummy-line variant, with a single dummy line separating each measurable line to create a range of 9.52µm with a pitch of 560nm:

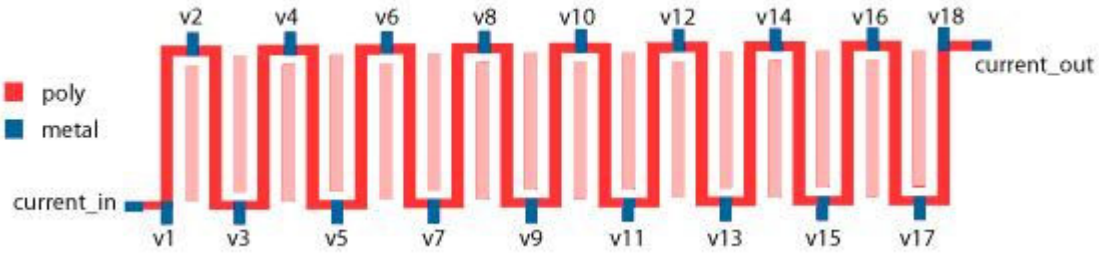

Figure 3.14. Micron-scale ELM structure with single dummy line. This variant on the base-case ELM structure doubles the total range from leftmost measurable line to rightmost measurable line, while doubling the granularity of measurements (distance between successive measurable lines).

In additional variant structures, increasing numbers of dummy lines are inserted between the measurable polysilicon lines so that the total separation distance of measurable lines increases geometrically in powers of 2 (2, 4, 8, 16, and so on up to 256 times the minimum separation distance of 280nm), thereby expanding to a maximum separation range of 1.15mm, measured from one end of the 256-dummy-line test

structure to the other. Several of the additional variant structures are illustrated in **Fig. 3.15**.

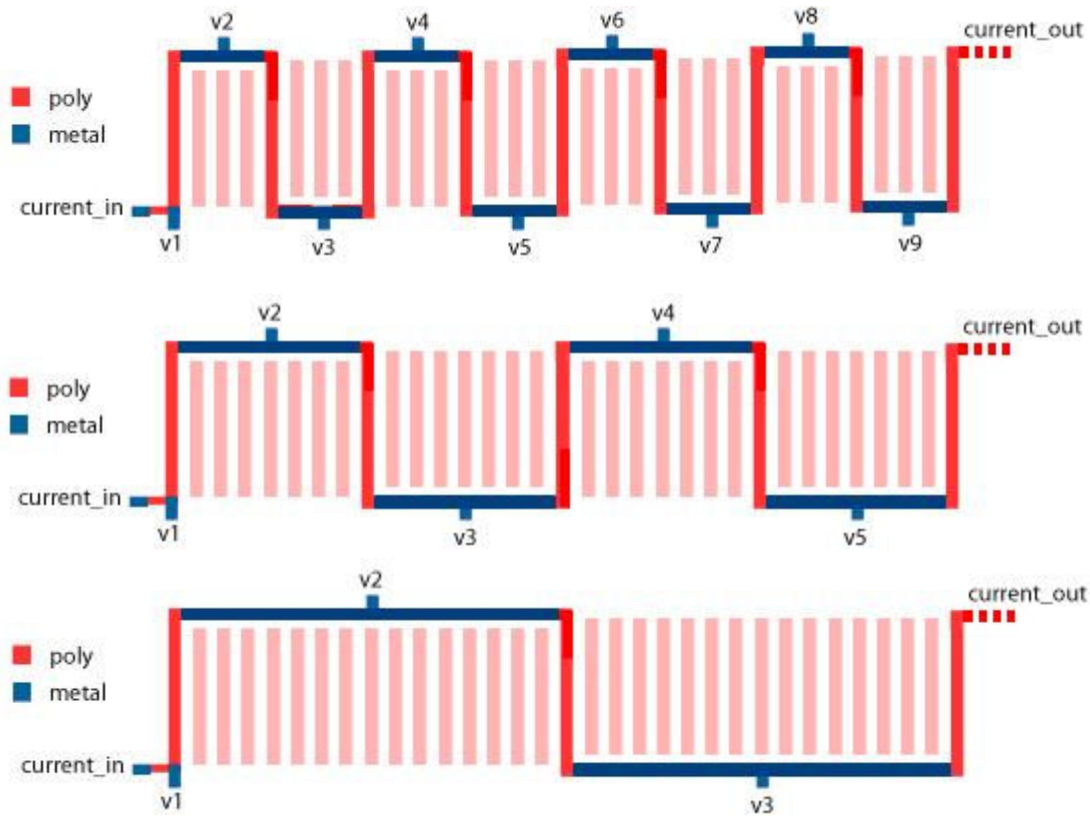

Figure 3.15. Micron-scale ELM structures with multiple dummy lines (3, 7, and 15 respectively). By maintaining the pitch of measurable lines and dummy lines, measurements from structures with varying number of dummy lines may be fairly combined into a single data set.

In sum, 9 total polysilicon structures (or DUTs—devices under test) were designed, with each structure contributing 17 total linewidth measurements. Each set of 9 DUTs was instantiated both in a horizontal and a vertical orientation on every test chip, and a total of 25 test chips were manufactured. Therefore, a total of 3825 measurable polysilicon lines were available for characterization in each orientation. A bird's-eye view of the Cadence layout of a single instantiation of the 9 different DUTs is shown in **Fig. 3.16**.

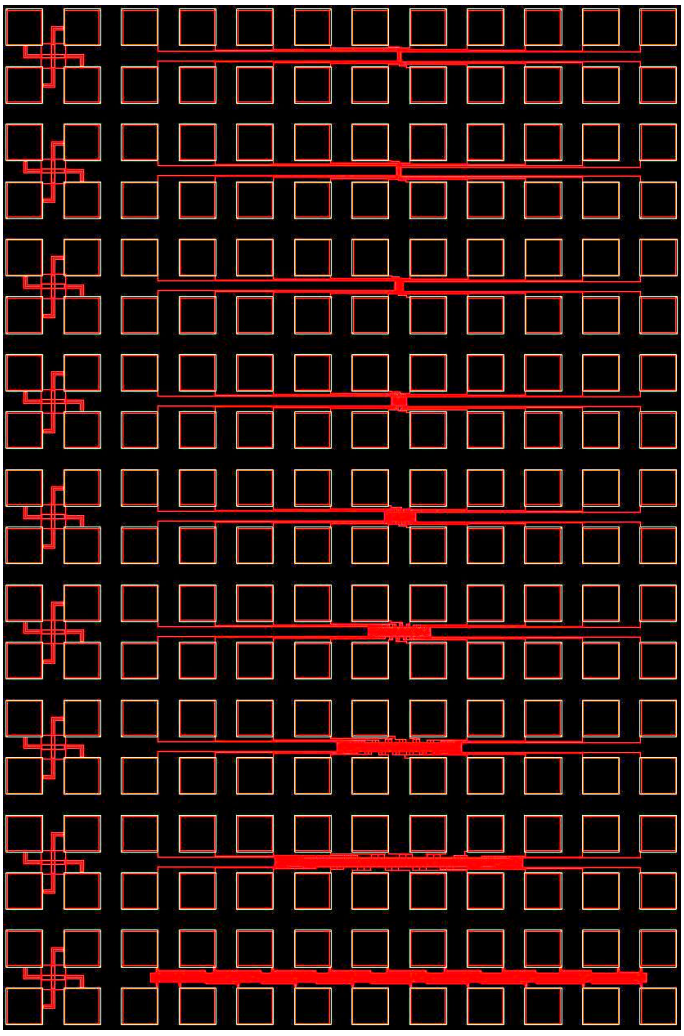

Figure 3.16. Global view of the Cadence layout of the test structure design. The nine test structures (increasing in range from top to bottom with increasing addition of dummy lines) are each connected to a frame of 2-by-10 metal pads for probing. Each test structure is accompanied by a Van der Pauw structure, located adjacently to the right.

At the device level, several design choices were made to assure that the measurements would capture the desired variation. In particular, the contacts from metal to polysilicon were designed to abut the precise edge of the line-under-test, so that the contact resistance would not contribute significantly to the resistance of the polysilicon line (**Fig. 3.17**).

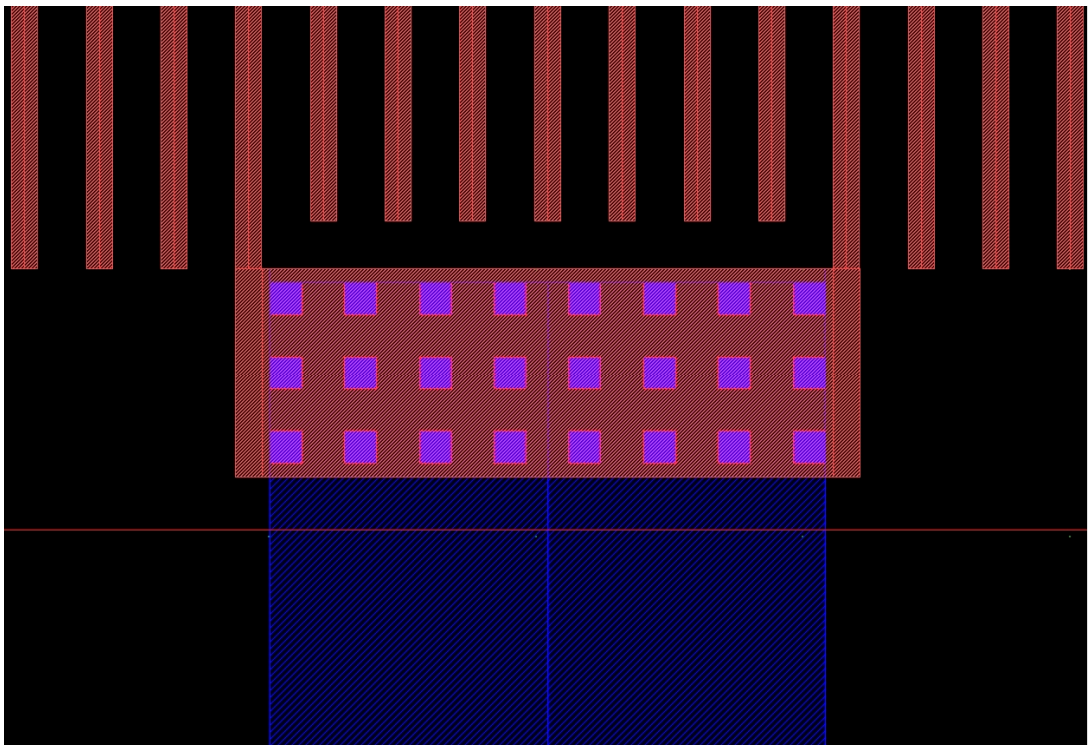

Figure 3.17. Snapshot of the Cadence layout of the metal-to-poly contact region of the micron-scale test structures. Contacts fill the region to the maximum allowed by design rules to mitigate contact resistance issues.

The success of this design feature was verified by comparing two Kelvin-type measurements: first, the resistance of two adjacent polysilicon lines was measured, and second, the resistance of the two lines in series was measured. By comparing the sum of resistances targeted in the first measurements with that captured in the second measurement, the contact resistance could be measured, as described in **Eq. 13-15**.

$$res_{\sin gle} = res_{line} + 2 * res_{contact} \tag{3.13}$$

$$res_{double} = 2 * res_{line} + 2 * res_{contact} \tag{3.14}$$

$$res_{\sin cle\_A} + res_{\sin gle\_B} - res_{double\_AB} = 2 * res_{contact} \tag{3.15}$$

The measured contact resistance was 0.6Ω on average, insignificant in comparison to the roughly 1700Ω resistance of the polysilicon line.

3.3.2 Measurements and Analysis

The test structures were measured using an Electroglas Automatic Probe Station. To increase measurement precision, several measurement iterations were performed at each ELM site. By averaging over several iterations, the precision was lowered to 0.1nm 3-σ. (This was established by estimating the variance of several sets of repeated measurements over the same structure, with the replication of the entire sequence including alignment and probing). In sum, roughly 30 seconds were required to test each measurable polysilicon line. Therefore, the required measurement time for the 7650 total measurements compiled from all available test structures was roughly 65 hours.

The data from the structures with horizontal orientation was analyzed separately from the data for the structures oriented vertically. Combining the 3825 measurements from each orientation, the total variance in CD was 7.63 nm^2 for the horizontal orientation, and 5.32 nm^2 for the vertical orientation. However, the majority of this variation was due to chip-to-chip variation (6.85 nm^2 for horizontal; 4.64 nm^2 for vertical), as shown in **Fig. 3.18**. This effect was expected: since the test chip (roughly 2mm by 4mm) only comprised a small fraction of the total lithographic field (likely 1.5cm by 1.5cm), the bulk of observed variance ought to be chip-to-chip in comparison to within-chip. The large bias between variation for the vertical orientation and horizontal orientation of these test structures could be attributed to a number of competing factors, including bias in the illumination settings or lens aberrations, as well as bias in the neighboring designs of the multi-project reticle. Unfortunately, the limited transparency of the foundry process restricted the scope of our investigation in this matter.

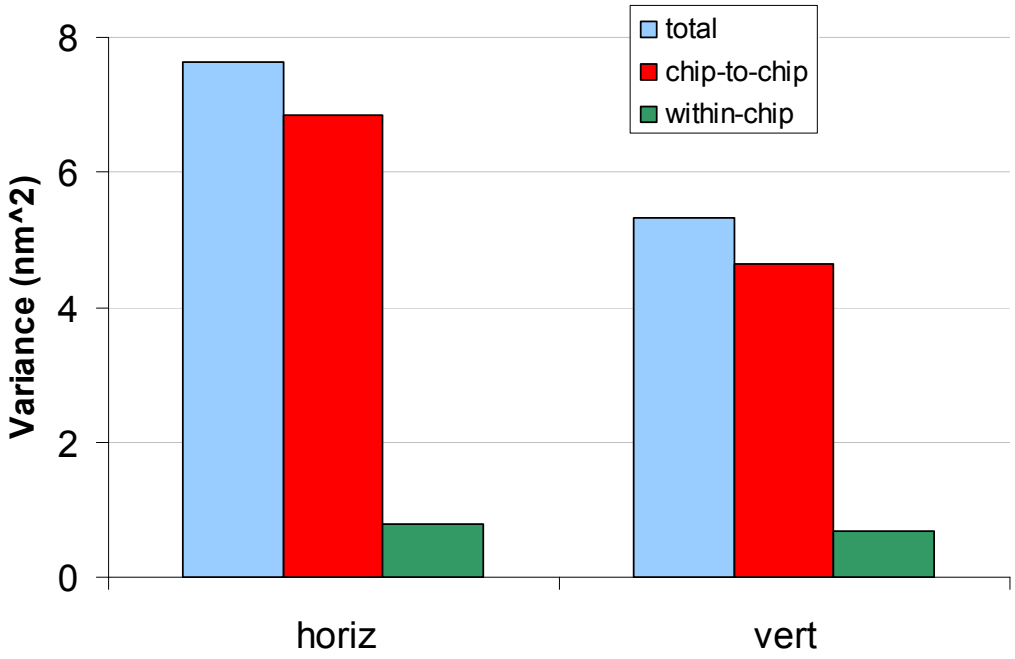

Figure 3.18. Comparison of total variance and chip-to-chip variance, by orientation. For each orientation, the total variance is decomposed into a chip-to-chip component and a within-chip component. The chip-to-chip component is much larger than the within-chip component, as expected since the chip size is a small fraction of the total lithographic field; the small within-chip component is the focal point of our analysis.

However, the goal of the experiment was to characterize the short-range CD variation, so the relatively small within-chip component of variation is our focus. To remove the chip-to-chip component of variation, the difference between the mean CD from each chip and the global mean CD across all chips was subtracted from each measurement within each chip. Then, by averaging across the 25 chips for each position (1-17) within each test structure, and for each test structure within each chip (1-9), we are able to generate a useful illustration of the raw measurement data, shown in **Fig. 3.19**.

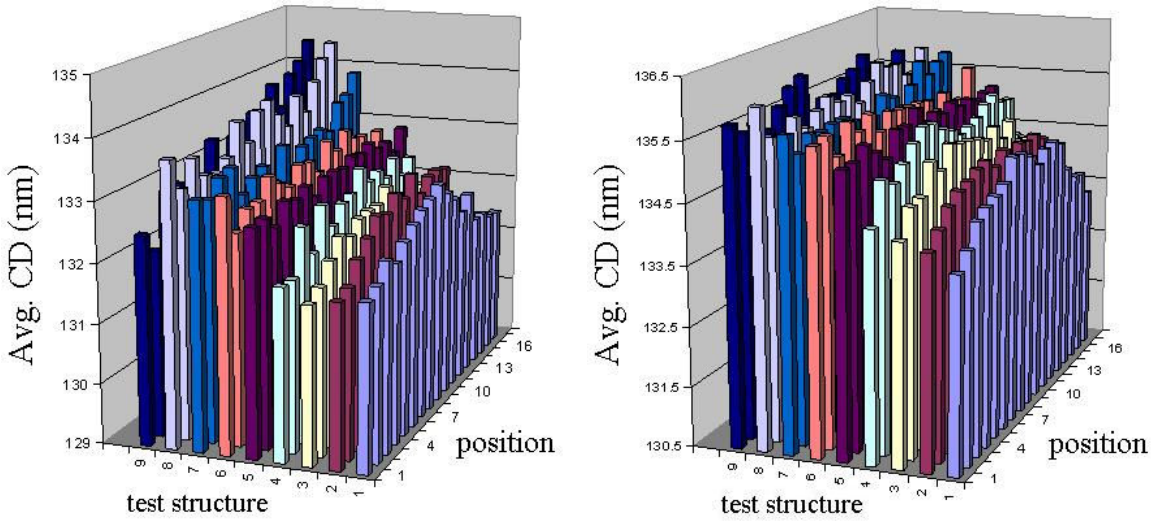

Figure 3.19. Normalized CD data for horizontal (left) and vertical (right) orientations. Due to the insertion of dummy lines, the intra-structure separation distance increases as we go from test structure 1 (zero dummy lines) to test structure 9 (256 dummy lines).

It is clearly evident from these plots that some additional deterministic variation exists within these normalized data; specifically, there appears to be a pattern-dependent effect for the test structures with relatively few dummy lines—structures 1 through 4. The values of critical dimension near the edges of these structures are consistently lower than the values of CD at the middle of the structure. Moreover, the average CD of the entire structure is less than the average CD of the structures with a larger number of dummy lines.

Two possible explanations for this source of deterministic variation were examined. First we looked at the potential impact of lithographic iso-dense bias. The lines at the edge of the structures with relatively few dummy lines would see a higher local density of light from the exposure tool since fewer neighboring lines would contribute intensity by optical proximity. To investigate this possibility, inspection was carried out using the Calibre lithography simulation tool [3.17]. Using the layout of one of the ELM test

structure sites, a section of the periodic serpentine polysilicon structure shown in **Fig. 3.20** was submitted to simulation of resist image.

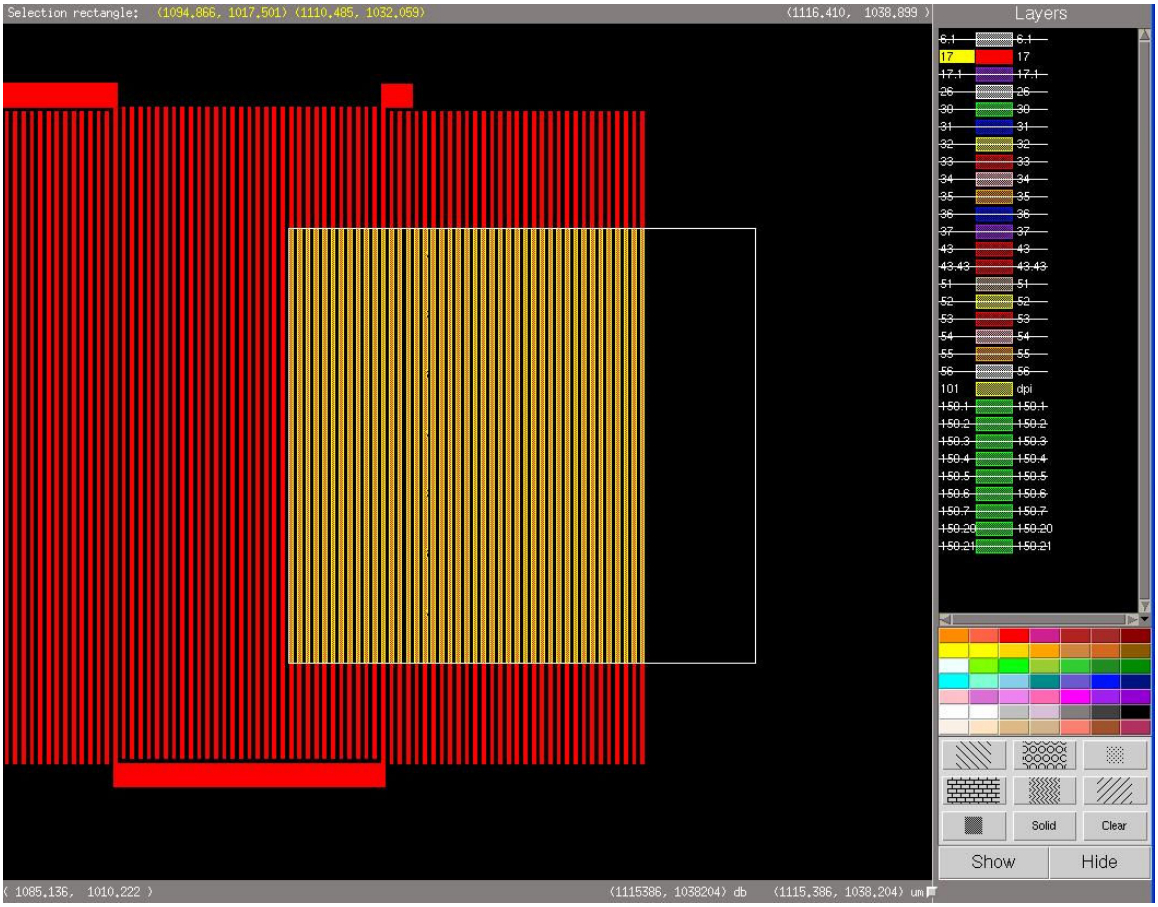

Figure 3.20. Snapshot of the lithography simulation carried out on the test structure using Calibre [3.17]. The area within the bounding box shaded a lighter color is the region that was subjected to rigorous resist image simulation.

The result is displayed in a more detailed view in **Fig. 3.21** with annotated linewidths of the rightmost four lines of the test structure. It is immediately evident that the measured linewidth of the serpentine pattern as used in this test structure achieves regularity within only a few wavelengths of the exposing light; the effect of optical proximity only extended 3 to 4 neighboring lines, as the first line has a simulated CD of 0.129µm, the second line has CD 0.134 µm, and the third, fourth, and rest of the interior

lines have a simulated CD of 0.136 μm. Since even the structures with no dummy lines were designed to have 5 lines on both exterior sides of the whole structure for lithographic uniformity, the observed pattern-dependent bias cannot be attributed to a lithographic effect.

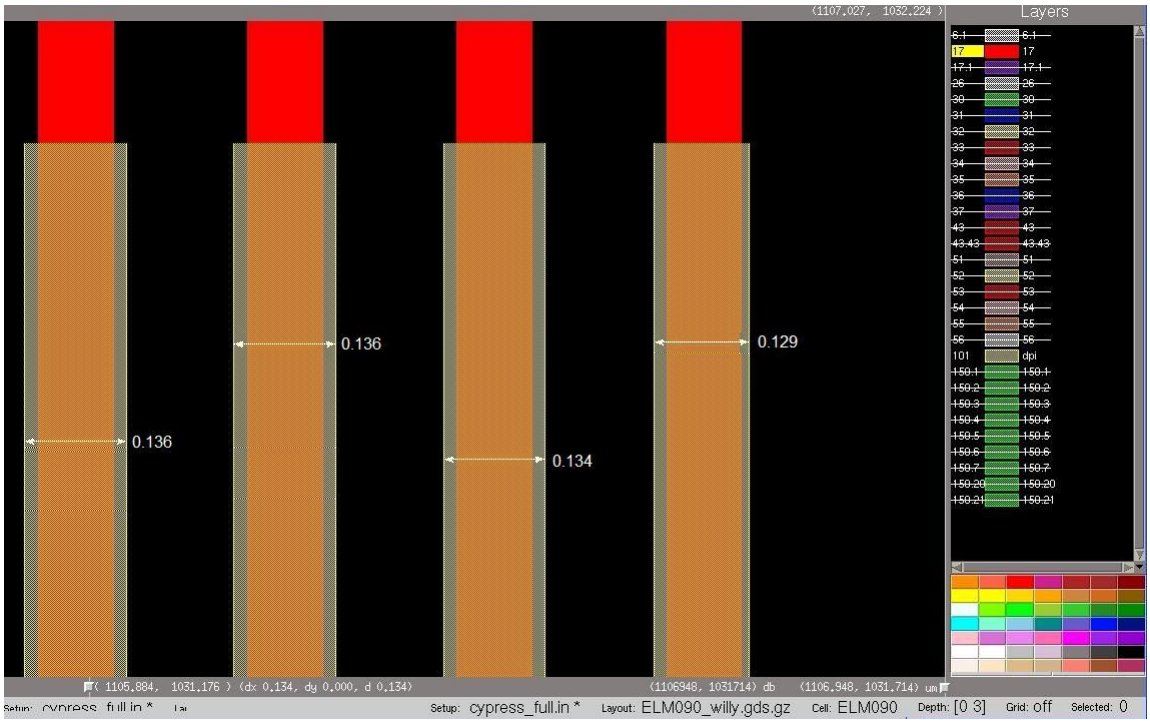

Figure 3.21. Close-up snapshot of the Calibre resist image simulation on the ELM test structure. The narrower lines at the top indicate the designed-for CD (on reticle); the wider, lighter lines that overlap the designed-for CD layer indicate the resulting simulated resist image. By the third line onward into the structure, the resist image linewidth achieves regularity. (Simulated resist image linewidths are annotated in units of microns.)

In turn, it is strongly suspected that the observed result is due to local loading effects in the etch process. In areas of higher local density, the lower availability of etch species as well as the limited rate of byproduct removal lead to a slower local etch rate. By contrast, the lines with lower local density are etched more quickly, since they have greater "exposure" to the etching process. To capture both the tapering at the edges of

the low-density structures as well as the lower overall CD for the structures with fewer total lines, a two-term etch loading model was proposed. The first term, capturing micron-scale localized loading, uses a narrow moving window centered on each successive line within each structure to determine the local density as seen by that particular line. The width of the window was tuned to ~2 microns to maximize the model's predictive capability. An illustration of this term of the model is shown in **Fig. 3.22**.

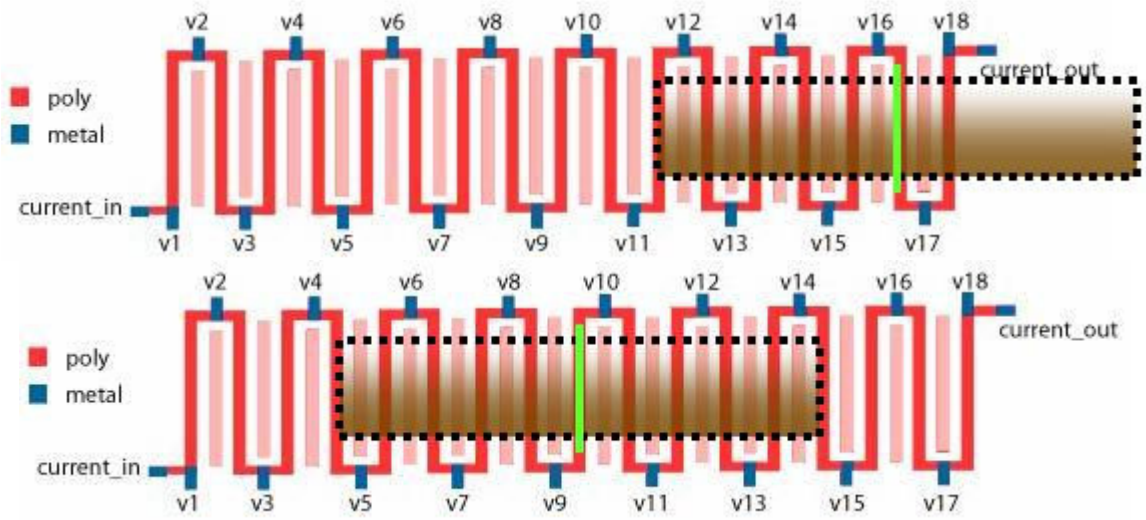

Figure 3.22. Illustration of line-by-line local density model. For each measurable line within the test structure, the local polysilicon density is given by the total polysilicon area falling within a 2-micron-wide window centered on the line in question. For lines in the center of a test structure, the local density is greater than for lines near the structure's edge.

The second term, capturing the structure-wide "global" loading, simply used the count of total polysilicon lines within the structure as the independent variable. This term is illustrated in **Fig. 3.23**.

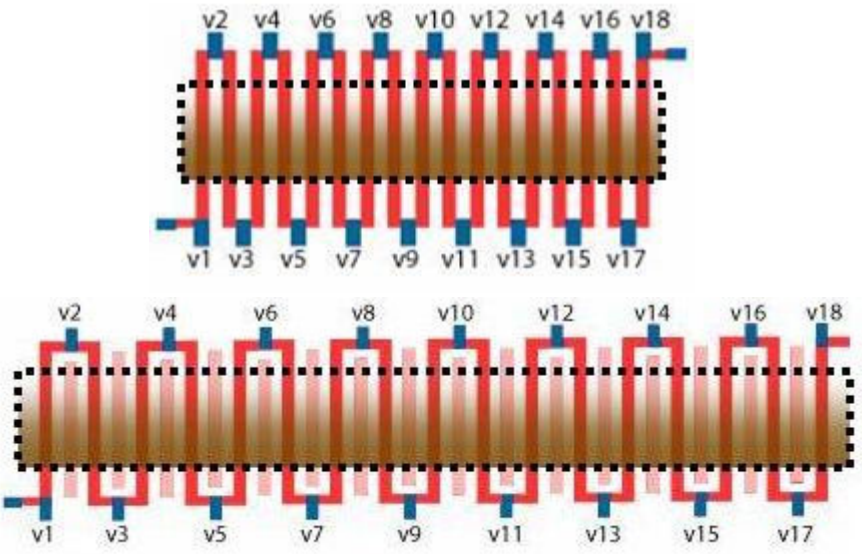

Figure 3.23. Illustration of structure-wide "global" density model. For each test structure, the global density is given by the total polysilicon area encompassed by the test structure. Those test structures with a large number of dummy lines will have a greater global density value than test structures with few or no dummy lines.

The two terms were combined into a single linear model and coefficients were estimated using least squares regression. The resulting models are:

$$CD_{i,horiz} = 129.2 + 0.1188 \times D_{i,Local} + 0.00019 \times D_{i,Structure} + \varepsilon_i \quad (3.16)$$

$$CD_{i,vert} = 131.17 + 0.1340 \times D_{i,Local} + 0.00014 \times D_{i,Structure} + \varepsilon_i \quad (3.17)$$

where $D_{i,Local}$ is the local density for the ith polysilicon line, and $D_{i,Structure}$ is the global density for the structure containing that line. This cumulative density model accounted for an additional 0.25 nm^2 (horizontal orientation) and 0.24 nm^2 (vertical orientation) of total variance in the dataset. Relevant measures of statistical significance are given in **Tables 3.1** and **3.2**.

Table 3.1. Parameter estimates for pattern-dependent model on horizontal CD variation.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	129.2030	0.119685	-29.89	<.0001
Local Density	0.1187828	0.004266	27.84	<.0001
Total Density	0.000191	9.139e-6	20.90	<.0001

Table 3.2. Parameter estimates for pattern-dependent model on vertical CD variation.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	131.1707	0.109402	-36.23	<.0001
Local Density	0.1339791	0.0039	34.36	<.0001
Total Density	0.0001421	8.353e-7	17.02	<.0001

In parallel, two alternative models for pattern-dependent plasma etch bias were considered. First, a model for etch rate as a negative exponential function of effective pattern density was considered [3.18], where the exponential coefficient was tuned to maximize the model's predictive model. Second, a specialized model of a (1/r) form was considered and again tuned to create the best fit to these data [3.19]. However, neither of these alternative models provided as much predictive accuracy as the proposed two-term moving window approach outlined above, as measured by R^2 of the model fit.

After applying the proposed model and calculating the residuals, we can see that the pattern-dependent model cannot account for all systematic effects in these data, since there still appears to be correlation by pattern group shown in **Fig. 3.24**. Again, these remaining systematic effects can be attributed to unknown variability in the surrounding pattern density of designs adjacent to our test structures on the multi-project test reticle, as well as to the limited predictive capability of the pattern-dependent etch bias model used here. Unfortunately, since information about neighboring designs is unavailable, these remaining effects can only be acknowledged as systematic but not modeled.

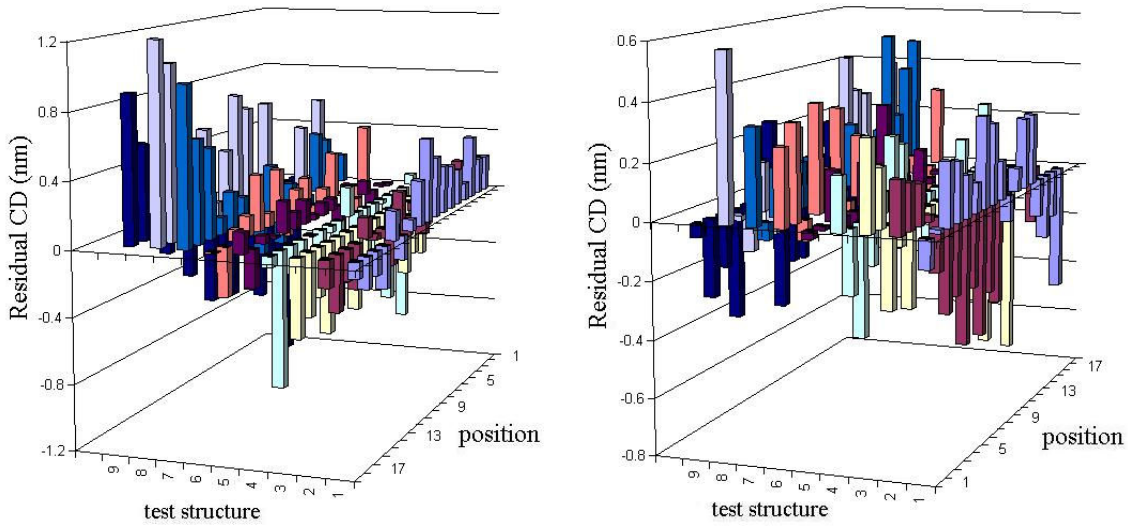

Figure 3.24. Residual CDs for horizontal (left) and vertical (right) orientations. Some slight systematic clustering effects can be seen, and are attributed to slight inaccuracy in the pattern-dependent etch bias model as well as to variations in polysilicon density of adjacent circuits on the multi-project reticle.

A breakdown of the within-chip variance is shown in **Fig. 3.25**. It is evident that the modeled pattern-dependent deterministic variation comprises roughly one-third of the total within-chip variance. Finally, spatial correlation analysis is performed on the stationary distribution of residual "random" variation left once all the deterministic components are removed. As shown in **Fig. 3.26**, the autocorrelation derived from the dataset is barely distinguishable from zero, for all separation distances.

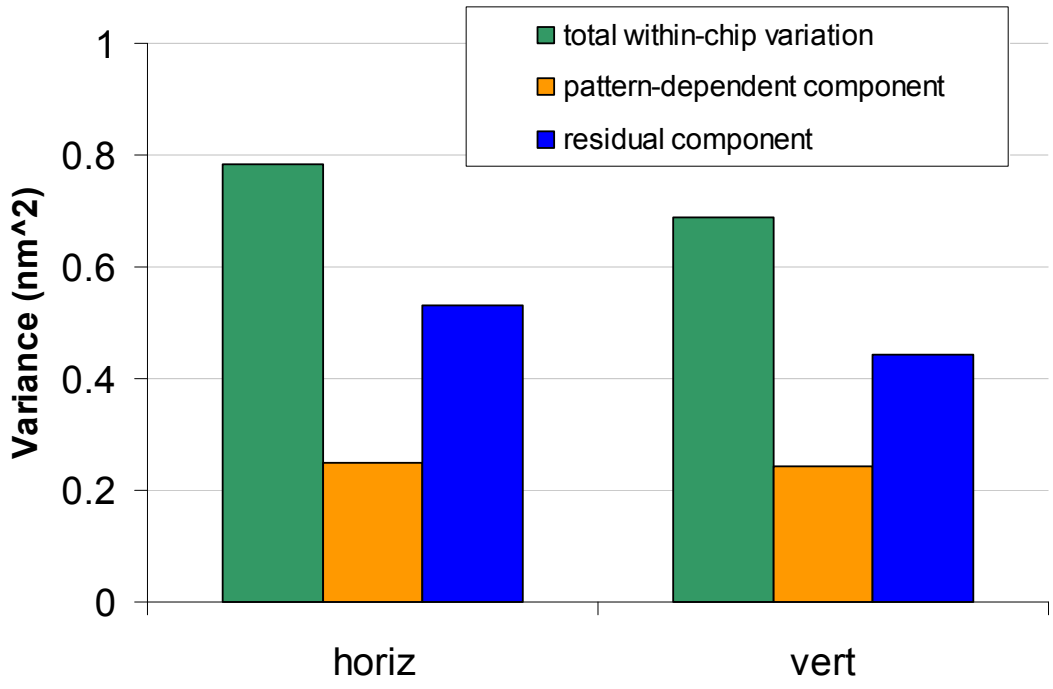

Figure 3.25. Decomposition of within-chip variance for both orientations. The pattern-dependent etch bias accounts for roughly one-third of the within-chip variance; the residual component contains both random variation and some small systematic components that could not be rigorously modeled.

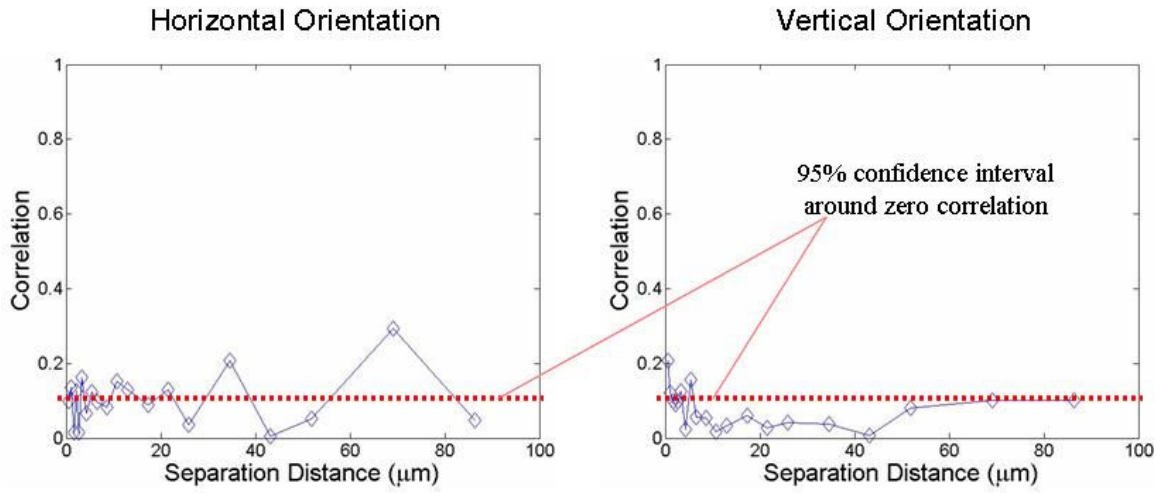

Figure 3.26. Spatial autocorrelation plots for the residual CD distribution generated after removing the chip-to-chip variation component as well as the pattern-dependent within-chip variation component.

For some separation distances, the spatial correlation falls outside of a 95% confidence interval around zero correlation, but the magnitude of correlation for those separation distances is still very small. It is likely that the very small amount of spatial correlation (consistently positive) still observed is indicative of the lack of perfect fit of the short-range pattern dependent model, as well as nonlinearity in the chip-to-chip variation that might be contaminating the residual CD distribution. Since some deterministic variation will not be ideally captured by modeling, some spatial correlation will be observed in the residual distribution. In this case, specifically, a clear component of deterministic spatial variation remains unexplained in DUT 9 for the horizontal orientation, where the CD measurements systematically increase from position 1 to position 17. This component was left unexplained because no assignable cause could be described with certainty. However, these test chips were fabricated as part of a multi-user foundry service where several designs are combined on a single reticle, and the characteristics of neighboring designs are not shared between users. Therefore, it is suspected that a gradient in polysilicon density in the area of the reticle immediately adjacent to DUT 9 might have played a role, since DUT 9 is located on one edge of this test chip. Since etch microloading effects have been shown to have relatively long ranges (in the millimeter scale) [3.18], such a gradient in polysilicon density might give rise to a similar gradient in CD measurements from DUT 9. In any case, as shown here, if a rigorous decomposition is performed, the remaining correlation will essentially be negligible.

References

[3.1] T. A. Brunner, C. P. Ausschnitt, "Process monitor gratings," *Metrology, Inspection, and Process Control for Microlithography XXI*, Proceedings of SPIE vol. **6518**, 2007.

[3.2] K. R. Lensing, et al, "Lithography equipment control using scatterometry metrology and semi-physical modeling," *Metrology, Inspection, and Process Control for Microlithography XXI*, Proceedings of SPIE vol. **6518**, 2007.

[3.3] Y. Lee, M. Sung, E. Lee, Y. Sohn, H. Bak, H. Oh, "Temperature Rising Effect of 193nm Chemically Amplified Resist during Post Exposure Bake," *Advances in Resist Technology and Processing XVII*, Proceedings of SPIE vol. **3999**, pp. 1000-1008, 2000.

[3.4] P. Friedberg, et al, "Time-based PEB adjustment for optimizing CD distributions," *Metrology, Inspection, and Process Control for Microlithography XVIII*, Proceedings of SPIE vol. **5375**, pp. 703-712, 2004.

[3.5] Q. Zhang, P. Friedberg, C. Tang, B. Singh, K. Poolla, C. Spanos, "Across-wafer CD Uniformity Enhancement through Control of Multi-zone PEB Profiles," *Metrology, Inspection, and Process Control for Microlithography XVIII*, Proceedings of SPIE vol. **5375**, pp. 276-286, 2004.

[3.6] M. D. Smith, C. A. Mack, J. S. Petersen, "Modeling the impact of thermal history during post exposure bake on the lithographic performance of chemically amplified resists," in *Advances in Resist Technology and Processing XVIII*, Proceedings of SPIE vol. **4345**, pp. 1013-1021, 2001.

[3.7] Q. Zhang, et al, "Comprehensive CD Uniformity Control across Lithography and Etch," *Metrology, Inspection, and Process Control for Microlithography XIX*, Proceedings of SPIE vol. **5752**, pp. 692-701, 2005.

[3.8] M. Janssen, R. de Kruif, T. Kiers, "Reticle processing induced proximity effects," *18th European Conference on Mask Technology for Integrated Circuits and Microcomponents*, Proceedings of SPIE vol. **4764**, pp. 82-94, 2002.

[3.9] J. van Schoot, et al, "CD uniformity improvement by active scanner corrections," *Optical Microlithography XV*, Proceedings of SPIE vol. **4691**, pp. 304-312, 2002.

[3.10] B. Smith, "Forbidden pitch or duty-free: revealing the causes of across-pitch imaging differences," *Optical Microlithography XVI*, Proceedings of SPIE vol. **5040**, pp. 299-407, 2003.

[3.11] C. Hedlund, H.-O. Blom, S. Berg, "Microloading effect in reactive ion etching," *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. **12**, issue **4**, pp. 1962-1965, 1994.

[3.12] R. Chang, Y. Cao, C. Spanos, "Dishing-Radius Model of Copper CMP Dishing Effects," *IEEE Transactions on Semiconductor Manufacturing*, vol. **18**, no. **2**, pp. 297-303, 2005.

[3.13] I. Ahsan, et al, "RTA-Driven Intra-Die Variations in Stage Delay, and Parametric Sensitivities for 65nm Technology," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, pp. 170-171, 2006.

[3.14] M. Pelgrom, A. Duinmaijer, A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. **24**, no. **1**, pp. 1433-1439, Oct. 1989.

[3.15] J. Cain, C. Spanos, "Electrical linewidth metrology for systematic CD variation characterization and causal analysis," *Metrology, Inspection, and Process Control for Microlithography XVII*, Proceedings of SPIE vol. **5038**, pp. 350-361, 2003.

[3.16] Lake Shore 7500/9500 Series Hall System User's Manual, Appendix A, "Hall Effect Measurements," pp. A.16-A.17.
http://www.lakeshore.com/pdf_files/systems/Hall_Data_Sheets/A_Hall.pdf.

[3.17] Calibre nmDRC,
http://www.mentor.com/products/ic_nanometer_design/mask_syn/calibre_drc/index.cfm.

[3.18] K. Abrokwah, P. Chidambaram, and D. Boning, "Pattern Based Prediction for Plasma Etch," *IEEE Transactions on Semiconductor Manufacturing*, vol. **20**, no. **2**, pp. 77-86, 2007.

[3.19] K. Okada, H. Onodera, and K. Tamaru, "Layout Dependent Matching Analysis of CMOS Circuits," *Analog Integrated Circuits and Signal Processing*, vol. **25**, pp. 309-318, 2000.

Chapter 4

Traditional SPICE-Based Monte Carlo Circuit Simulation Frameworks

To investigate the impact of CD variation on circuit performance, the thorough characterization of CD variation must be put to work in a Monte Carlo simulation framework. This chapter presents an overview of traditional, SPICE-based Monte Carlo simulation frameworks that are widely used to evaluate the impact of variation on circuit performance variability, and that were initially used in this work. First, a custom-made SPICE-based simulation framework is designed to determine the impact of a simplified model for CD variation in the circuit performance variability space. Next, a commercially available tool is used to search for design styles that might exhibit a reduced sensitivity to device variation. In both instances, the importance of accuracy in the statistical description of process variation will be highlighted.

4.1 SPICE-Based Simulation Framework Construction

Monte Carlo simulation frameworks have three main components: (1) the selection of a canonical circuit, on which the simulations will be run under varying assignments of device parameters, (2) the random instantiation of those device parameters, as specified by the model of variation, and (3) the simulation engine, which takes the specified canonical circuit and returns the desired simulation output. Usually, the canonical circuit is a simple, generic representation of a typical critical path, which is used in lieu of an actual design for simplicity and speed of simulation. For example, one might use a canonical circuit consisting of 10 equally-sized NAND stages, each loaded by a fan-out of two and separated from the following stage by 100μm of local interconnect, as shown in **Fig. 4.1** [4.1].

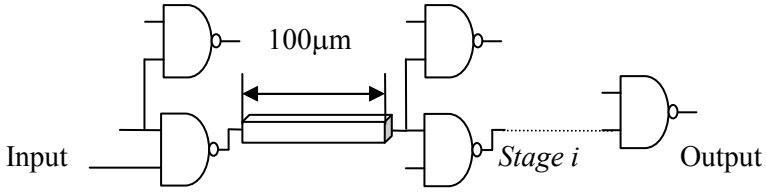

Figure 4.1. Canonical critical path circuit (10-stage NAND chain, Fan-Out = 2).

Alternatively, the canonical circuit may be tailored to target very specific areas of interest in circuit performance. For example, the critical path of a Kogge-Stone carry lookahead adder was cast into a canonical circuit to allow for comparison of variability under competing implementations [4.2]. The Radix 2 implementation (**Fig. 4.2**), with shorter stack height, smaller node fanout, and larger logic depth, was shown to have slightly less sensitivity to device parameter variation than the Radix 4 implementation

(**Fig. 4.3**). This result matches the intuition about the vulnerability of a critical path to variation set up in the previous chapter. Specifically, for greater the logic depth more averaging takes place when the sum delay or power consumption is calculated. Detailed results from this study will be presented in Section 4.3.

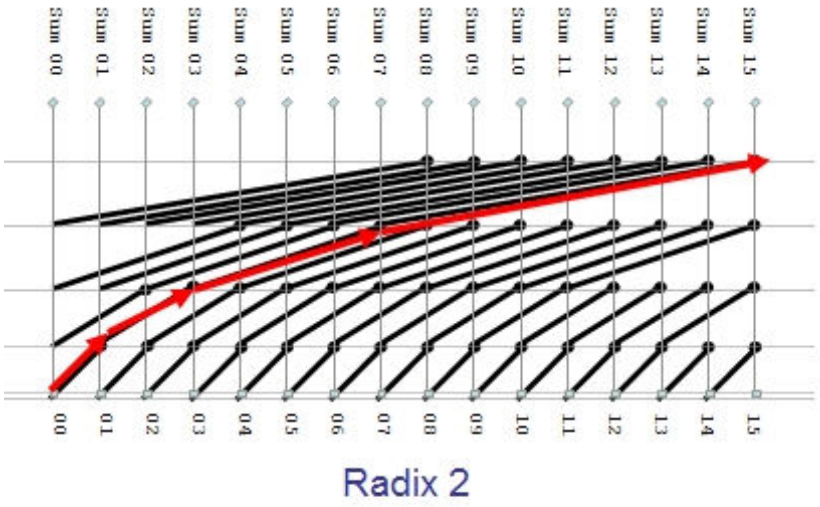

Figure 4.2. Canonical critical path diagram for Radix 2 Kogge Stone CLA.

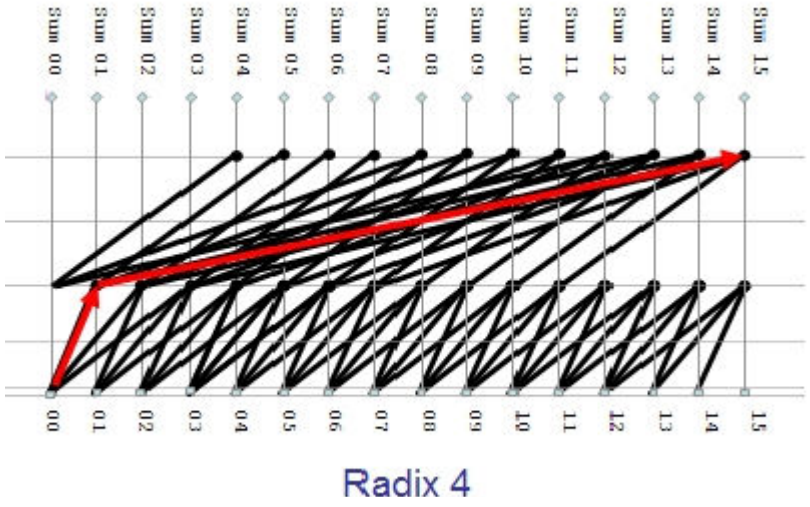

Figure 4.3. Canonical critical path diagram for Radix 2 and Radix 4 Kogge Stone CLA.

For the description of device variation, a wide range of models are used. Usually levels of variation are simply taken directly from the ITRS specifications [4.3]; rarely but occasionally they are tuned using measurement data from a specific production line. In this work, both the standardized, ITRS/BSIM variation levels [4.3, 4.5] and our more detailed characterization of CD variation from Chapter 3 will be implemented in Monte Carlo simulation frameworks to demonstrate the added value in using our more complex, spatially-aware model.

Finally, traditional Monte Carlo simulation frameworks typically rely on a circuit simulation software tool (such as SPICE—Simulation Program with Integrated Circuit Emphasis [4.4]) as the simulation engine. SPICE and other similar simulation programs take as their input a text netlist describing the circuit elements (in this case, the canonical circuit) and translate this description into a set of equations to be solved to determine the circuit performance. Since SPICE is equipped with a set of finely tuned device models [4.5] represented in library form, there is a relatively easy way to implement device variation; in each simulation run, certain device parameters can simply be overwritten, tweaking the SPICE models to reflect changes in sensitivities due to the randomly drawn device parameter values. By using a transient analysis within SPICE, a time series of voltages at various nodes may be collected, and from this, the delay of the canonical circuit may be calculated for each run.

4.2 Evaluation of a Simplified CD Variation Model

In previous work [4.1], a SPICE-based Monte Carlo simulation framework was used to characterize the impact of spatial variation on circuit performance variability, using a

simplified model of CD variation that relies solely on spatial correlation derived for the long-range CD dataset presented in Chapter 3. In **Fig. 3.5**, two spatial correlation characteristics were plotted; one captured the spatial autocorrelation in the residual CD variation distribution once all the deterministic components of variation had been removed, and the other captured the autocorrelation function calculated from the original data. The autocorrelation function calculated from the original data exhibited some counterintuitive features that arose due to violations of statistical assumptions, namely that the original CD distribution is not stationary. However, as a first pass at learning the impact of spatial CD variation on circuit performance variability, this CD variation model was used in spite of these violations. For further ease in selecting CD values in the Monte Carlo simulation framework, a rudimentary piecewise linear model was imposed to more simply capture the spatial correlation relationship. Illustrated in **Fig. 4.4**, the model contains two parameters: X_L, a characteristic correlation length, and ρ_B, the characteristic correlation baseline.

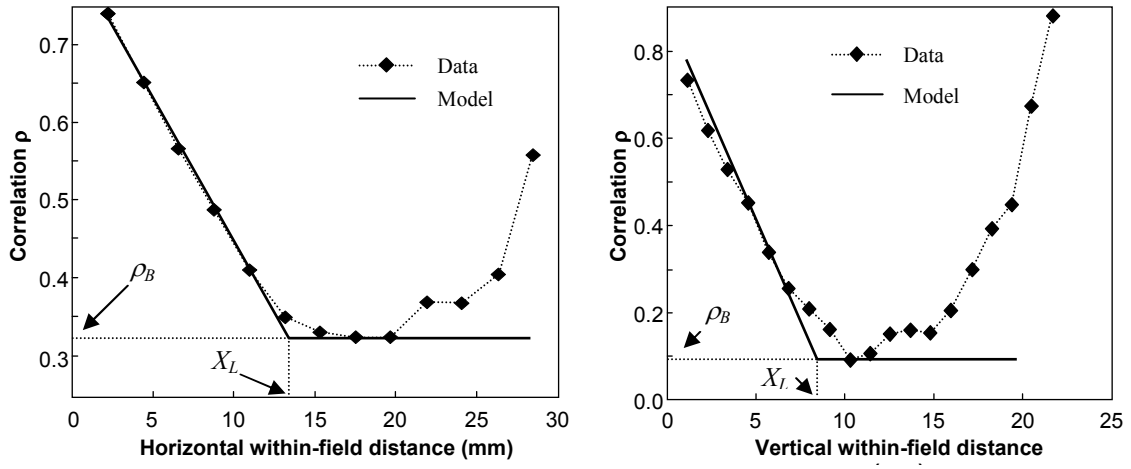

Figure 4.4. Spatial correlation dependences and models (horizontal on left, vertical on right). These piecewise linear models were used for ease of implementation in the Monte Carlo framework.

By definition, this model approximates the value of spatial correlation (ρ) given the separation distance (x or y) by:

$$\rho = \begin{cases} 1 - \dfrac{x}{X_L}(1-\rho_B), x \le X_L \\ \rho_B, x \ge X_L \end{cases} \tag{4.1}$$

Since these data are taken from a relatively large test chip covering the entire lithographic field (28 by 22 mm), values of separation distance greater than half the width of the field were excluded when fitting the model under the premise that critical paths would rarely span more than 1mm.

In that work, it was assumed that the spatial correlation in CD variation could be tuned independently from the CD variation itself. Under this assumption, there were three independent variables to consider for each Monte Carlo simulation run: X_L, ρ_B, and the normalized level of total gate length variation σ_L/μ_L. The NAND-2-chain canonical circuit displayed in **Fig. 4.1** was subjected to 3000 Monte Carlo simulations at 60 different combinations of X_L, ρ_B, and σ_L/μ_L, with nominal conditions of X_L = 1mm, ρ_B = 0.2, and σ_L/μ_L = 10%. Interconnect parameters and extracted RC parasitics were projected using the Berkeley Predictive Technology Model (BPTM [4.6]), and nominal device parameter values and variances were based on an industrial 90nm technology node.

The results, shown in **Fig. 4.5**, indicated that delay variability is most sensitive to the tuning of σ_L/μ_L; however, reduction of both X_L and ρ_B does display some benefit in terms of reducing delay variability. While the improvement due to direct reduction of parameter variation is linear, the improvement due to reduction of spatial correlation

parameter X_L increases gradually as the scaling factor is reduced, achieving over half of the total potential improvement with the final 25% reduction of X_L. Also notable is the fact that the greatest gains in reducing ρ_B are seen when X_L is also strongly reduced.

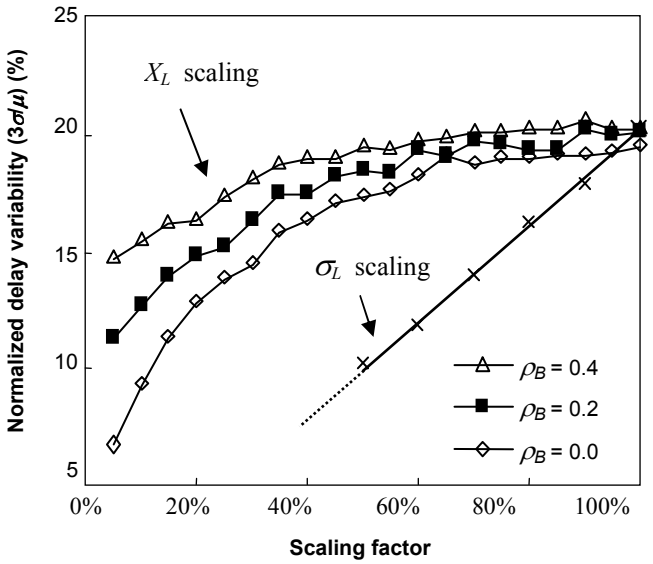

Figure 4.5. Comparison of impact of spatial correlation and normalized variation in gate length on critical path delay variability. Reducing levels of spatial correlation leads to a nonlinear reduction in delay variability; removing the last traces of correlation leads to the strongest marginal reduction.

However, it is important to keep in mind that this treatment of spatial correlation is an extremely rudimentary one for purposes of illustrating the use of a SPICE-based Monte Carlo simulation framework. Once the research group arrived at the realization that statistical assumptions were being violated in a way that might cause significant inaccuracy in predicting circuit performance variability, a new simulation framework was devised to correct the problem. This Monte Carlo framework would use a more thorough and accurate treatment of CD variation, including full decomposition of systematic components of variation, and will be presented in Chapter 5.

The execution time requirements for Monte Carlo simulations using this framework were burdensome; each single SPICE execution lasted roughly 10 seconds, meaning that

the full set of 180,000 total simulations took roughly one week to execute—using three parallel processors. This made the prospect of full precision Monte Carlo simulation less than practical.

4.3 Evaluation of Logic Implementation Styles

An additional study was carried out to use a commercially available SPICE-based Monte Carlo simulation framework for exploring circuit performance sensitivities to device variation under a range of various logic implementation styles. The tradeoffs between various logic styles in terms of nominal speed and power consumption were well-documented; however, relative levels of susceptibility to variation were less clear. Therefore, Monte Carlo analysis was used to assess the robustness of static CMOS, pulsed-static CMOS [4.7], DOMINO [4.8], and passgate-based LEAP [4.9] logic topologies to device variation.

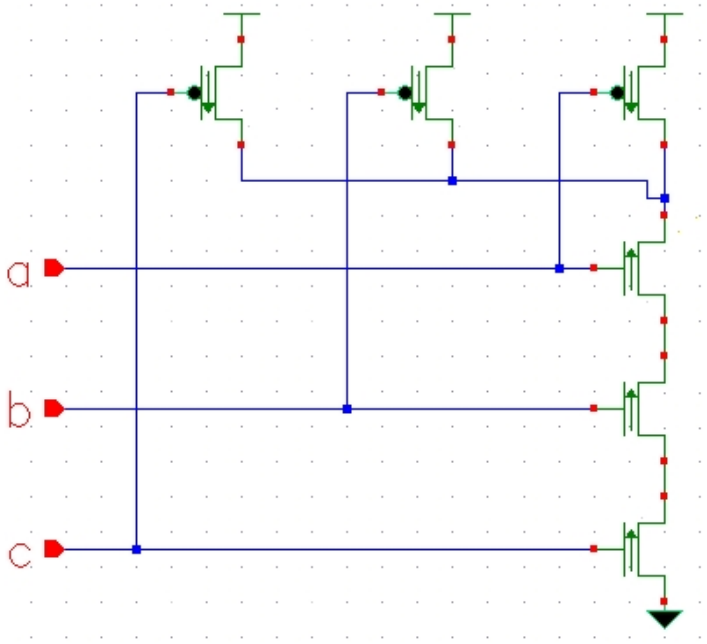

Figure 4.6. Static CMOS implementation of a NAND-3 stage.

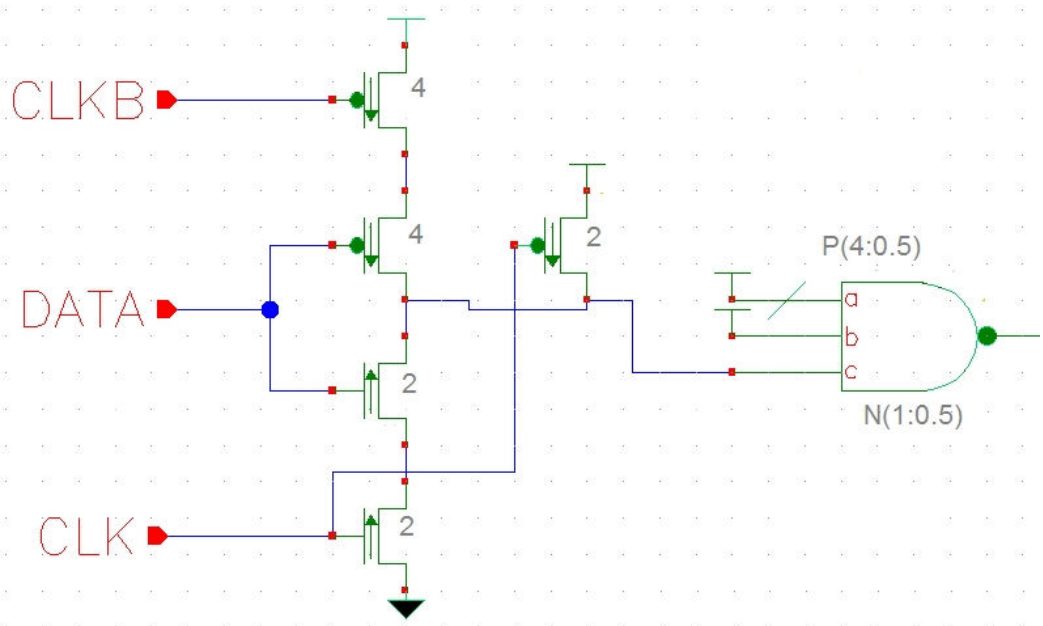

Figure 4.7. Pulsed-static CMOS implementation of a NAND-3 stage.

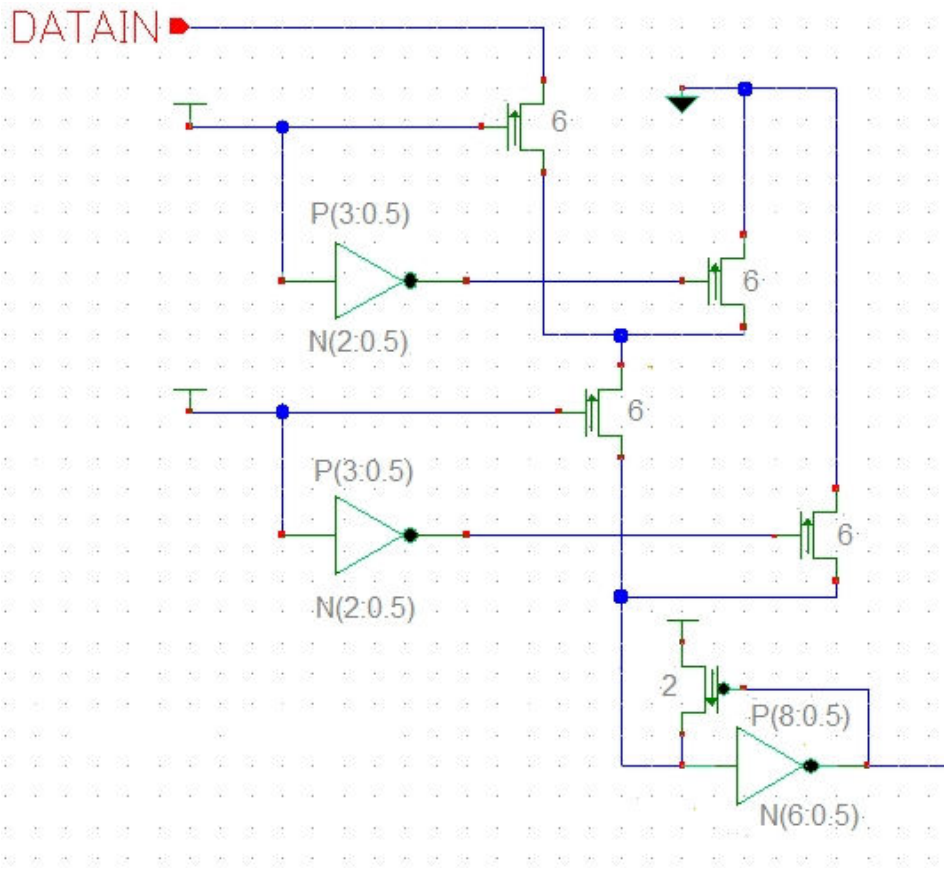

Figure 4.8. Passgate-based LEAP implementation of a NAND-3 stage.

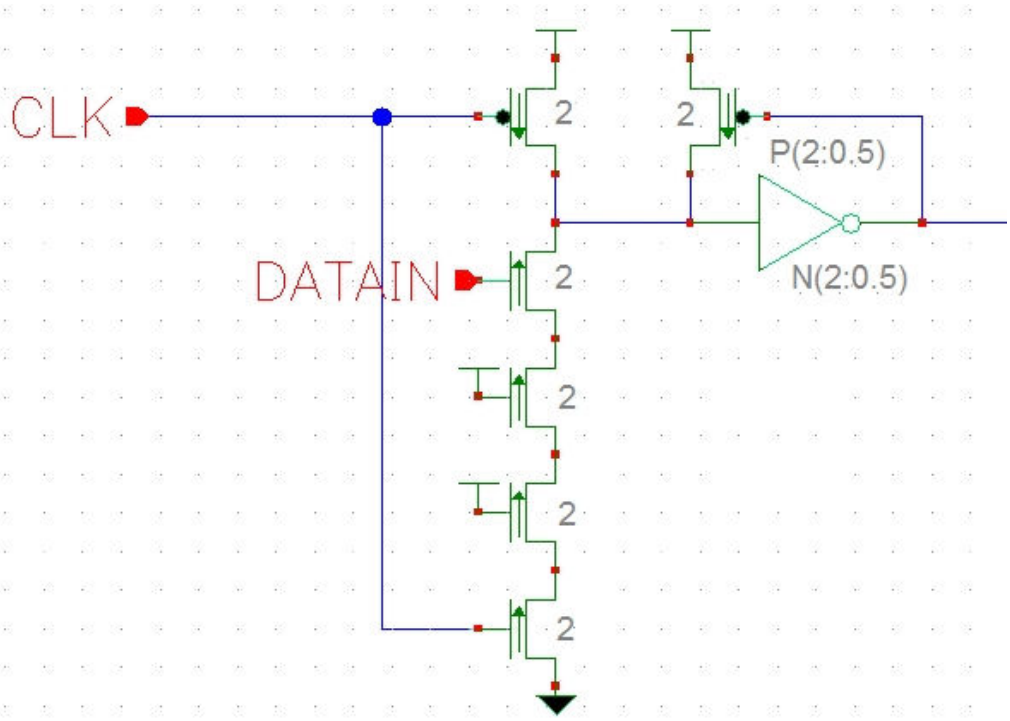

Figure 4.9. Dynamic DOMINO logic implementation of a NAND-3 stage.

An analysis was carried out for two canonical circuits that are representative of microprocessor datapath elements: a six-stage NAND chain and a 16-bit adder of various architectures. The device variation models for gate width, gate length, threshold voltage, and oxide thickness were built into the simulation package (and were purportedly based on both the BSIM SOI model [4.5] and hardware-based tuning). The operating supply voltage V_{dd} was varied over a normal distribution with a nominal value of 1V and 3σ-variance of 50mV. The spatial correlation coefficient ρ was set to 1.0 (i.e. parameter values are the same for all devices in a given circuit instance). Each circuit instance was subjected to two sets of Monte Carlo simulations. In the first simulation, all five parameters were varied simultaneously to capture the full statistical distribution in overall delay and active power. In the second simulation, each parameter was isolated and varied

individually, while all others were held at their nominal values. All interdependencies between parameters (e.g. V_{th} dependence on L, W, t_{ox}) were reconciled within the simulator, as asserted by BSIM models.

4.3.1 NAND-3 Chain Results

The canonical NAND chain consists of six three-input gates, with all non-switching inputs tied to V_{dd}. The output of the NAND chain is loaded with a static capacitor of value C_L = 10fF, consistent with the input capacitance of a typical stage. This static load is modeled as an ideal, passive capacitor in the SPICE simulation; its value remains constant throughout each of the simulations and is unaffected by the random parameter selection process. In order to compare this scheme with one that models the fluctuating input capacitance of an active successive stage, a second loading condition was evaluated, using fanout-of-three (FO3) loading. Because the FO3 load contains active devices, each of its transistors is subjected to the same process parameter variations as the other gates that form the chain. For each simulation set, 1000 SPICE simulations were completed.

As shown in **Fig. 4.10**, there were only slight differences in sensitivity to variation under the various logic topologies in terms of circuit delay. The passgate-based LEAP family appeared to fare the worst in the face of device variation for both loading schemes, but as indicated by the 95% confidence intervals surrounding each variation level (denoted by the red bars), the difference was slight in a statistical sense. As shown in **Fig. 4.11**, the passgate logic implementation also fared worst in terms of normalized power variability.

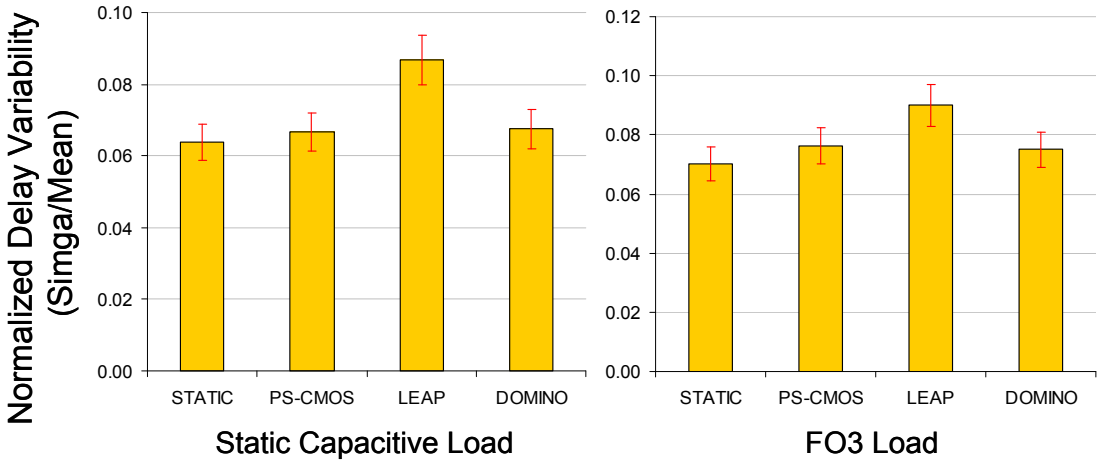

Figure 4.10 Nominal variability of a NAND chain under different loading schemes. For both loading schemes, the passgate-based LEAP logic style displays the greatest variability in delay. However, as indicated by the 95% confidence intervals included at the top of each bar, variability under different design styles displayed only slight differences.

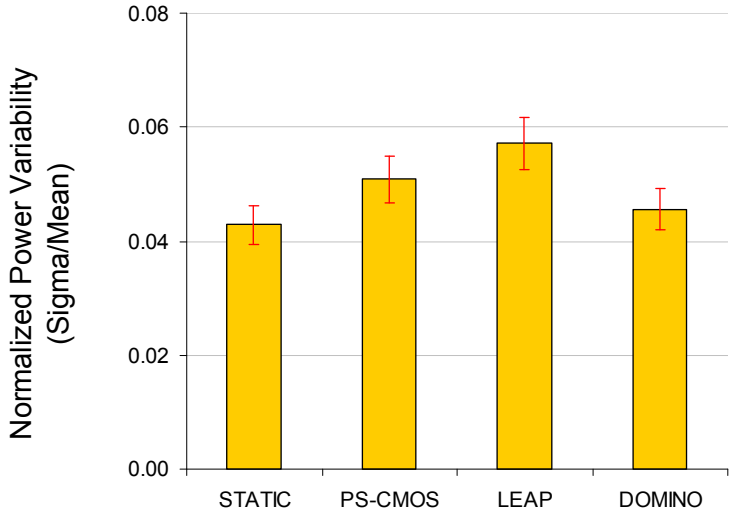

Figure 4.11. Normalized power variability of NAND chain with fixed capacitive load. Again, the passgate design style shows the largest variability in power; however, differences among the design styles are relatively small.

Fig. 4.12 shows the normalized individual parameter contributions to delay variability for the NAND chain with FO3 loads. Holding all other device parameters constant, each targeted device parameter was then allowed to take on randomly drawn values as in

previous simulations. Threshold voltage was identified as the most significant parameter in both cases, with an average variability contribution of 4.3%. Notably, the design that was most sensitive to variations in threshold voltage was the passgate-based LEAP implementation. Variation in gate length was nearly as significant as V_{th} contributions, accounting for an average of 3% of the overall variability. Furthermore, supply voltage variations accounted for average contributions of 2.4% (NAND chains). Finally, the process parameters t_{ox} and W were typically very well-controlled, and contribute on average 1.4% and 0.3% for the NAND chains.

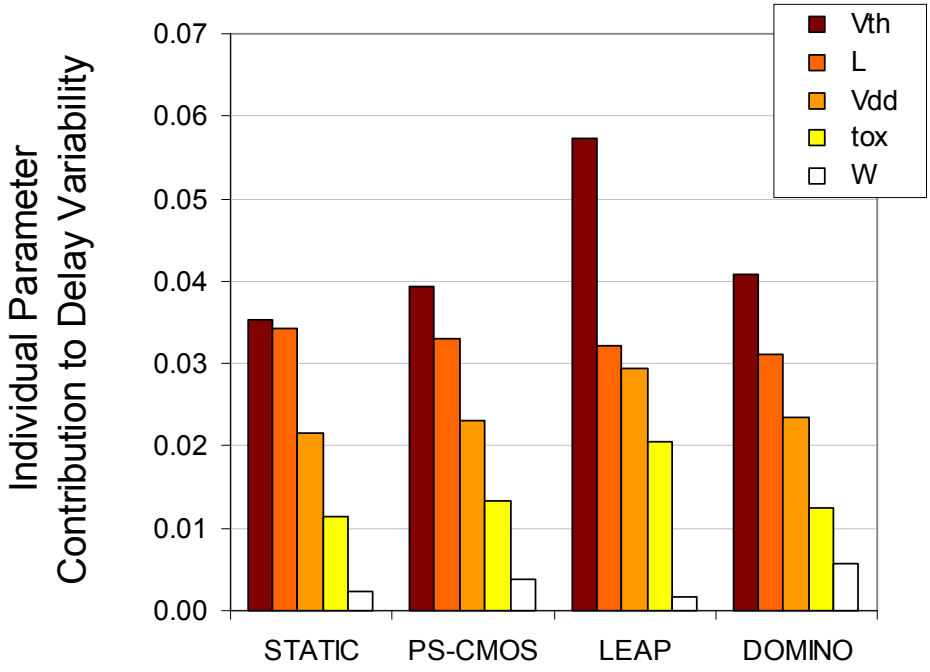

Figure 4.12. Individual parameter contributions to delay variability of NAND chain with FO3 loading (in units of sigma/mean).

These results quantify the high sensitivity of delay to fluctuations in V_{th}, V_{dd} and L across all logic evaluation styles. However, it is important to note that these particular results are heavily influenced by the specific levels of parameter variation specified in the commercial Monte Carlo framework tool that was tuned to a specific manufacturing line.

In particular, since gate length variation affects the threshold voltage of a transistor, it is difficult to separate the effects of purely random dopant fluctuation from gate length variation. In this case, the industrial partner seems to have assigned a significant portion of the overall threshold voltage variation to random dopant fluctuation, but that claim was not directly verified in this work.

4.3.2 16-bit Adder Results

In total, eleven 16-bit adders spanning a range of circuit architectures and logic evaluation styles were designed and submitted to both sets of Monte Carlo simulations. The three basic architectures are: ripple carry adder with a passgate-based Manchester carry chain (static and dynamic) [4.9], logarithmic carry-select (static, dynamic, and passgate) [4.9] and carry lookahead (Kogge-Stone radix 2 and radix 4 [4.10], Han-Carlson [4.11], and Brent-Kung [4.12]). A fanout-of-four (FO4) static inverter load was tied to the end of the critical path for all adder designs. Due to the increase in both logic complexity and transistor count as compared to the NAND chains, a reduced number of simulations (N = 200) was executed for each Monte Carlo sample in the case of the adders. Although this sample size was substantially smaller than for the NAND chains, it was sufficiently large to reveal insights into the performance variability for various implementations.

Simulation results for the family of 16-bit adders indicated that the static implementation of the carry-select adder displayed the smallest variation in delay (5.4%), as shown in **Fig. 4.13**. While variability levels for most other static and dynamic designs fall within 20% of the static carry-select, the passgate families again suffered the greatest

levels of variability. The three designs with the highest relative delay variabilities were the static ripple carry adder with passgate-based Manchester carry chain (7.1%), the passgate implementation of the carry-select (8.2%), and passgate-based radix 2 Kogge Stone (9.1%).

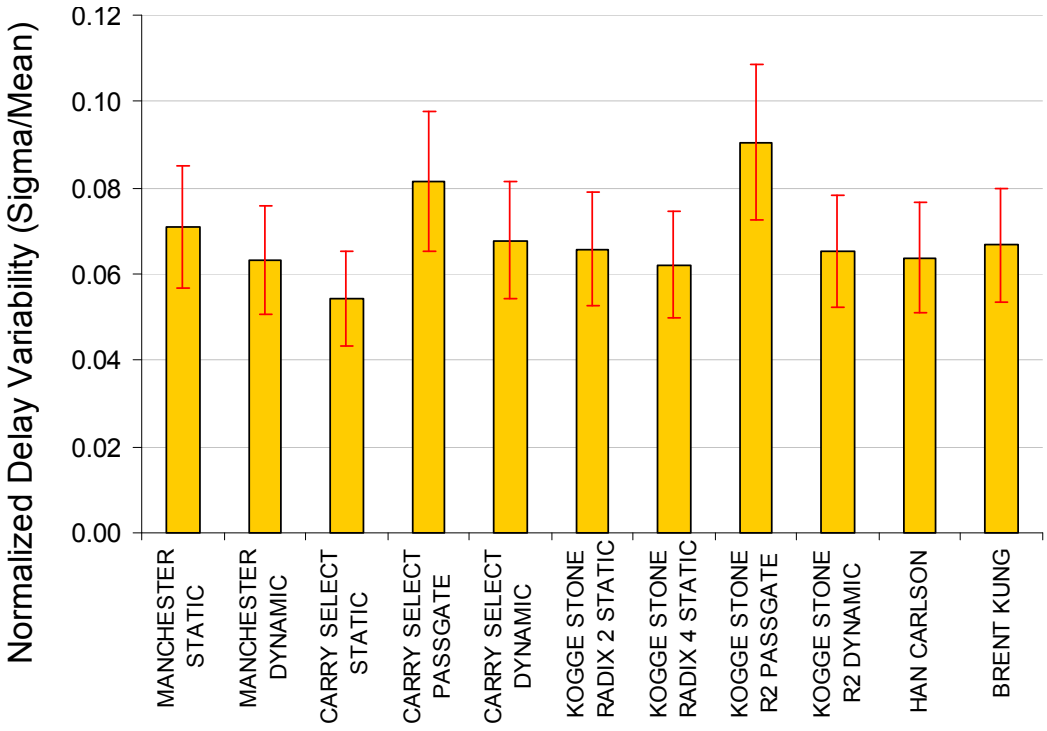

Figure 4.13. Normalized delay variability of 16-bit adders.

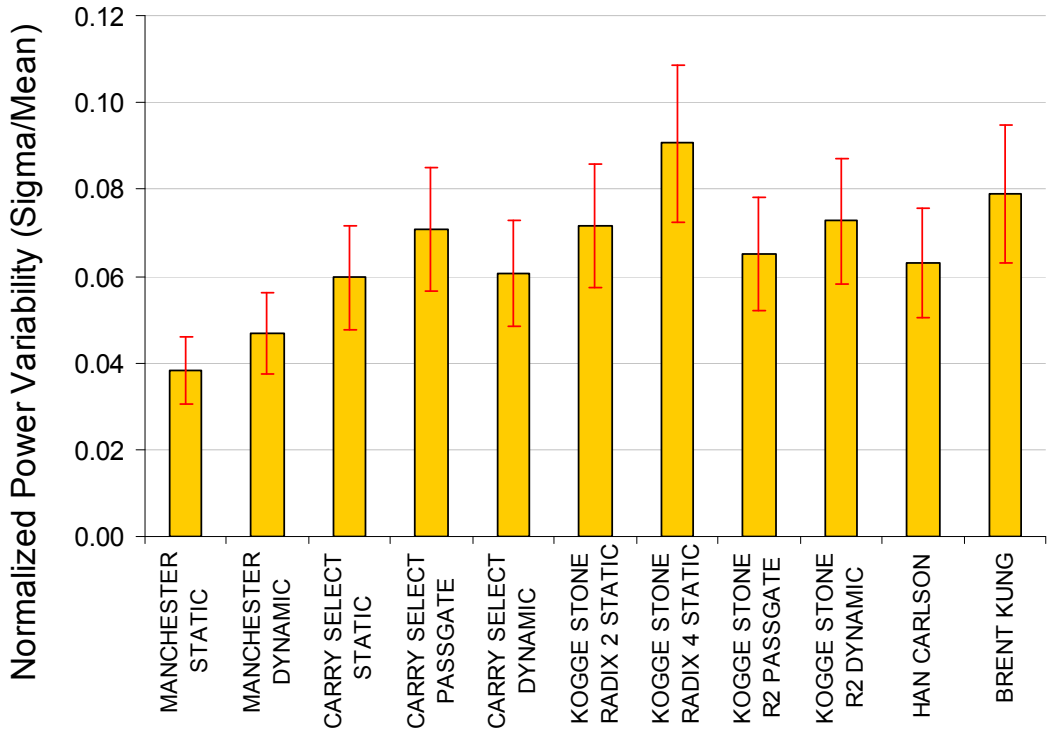

Figure 4.14. Normalized active power variability of 16-big adders.

Trends in adder power variability (active power only; leakage was not measured) are shown in **Fig. 4.14**. The static ripple carry adder using the Manchester carry chain displayed the most predictable power values (3.8% variability), while the variation in other designs range between 22% and 137% higher. The two least-robust designs from a power perspective were the static, radix 2 Brent Kung (7.9%) and static, radix 4 Kogge Stone (9.1%) adders, each with estimated variability twice as large as that of the Manchester carry chain implemented in static logic. This result may be attributed to the higher relative complexities of these designs, each having large intermediate capacitances along critical path nodes.

The Brent-Kung topology had widely varying internal fan-outs at each node, characteristic of its irregular tree structure, while the radix 4, Kogge Stone architecture had the tallest transistor stack height of all designs (four each of PMOS and NMOS).

These loads were composed of internal capacitances of active transistors, which fluctuate according to variations in process parameters. During adder operation, the active power drawn to continuously charge and discharge these varying capacitances fluctuates correspondingly, resulting in the higher relative power variability for these complex architectures. In comparison, the more regular adder architectures displayed less performance spread.

The normalized individual parameter contributions to delay variability for the adders are shown in **Fig. 4.15**. The conclusions were much the same as those for the NAND chain; namely, the threshold voltage was the single largest contributor to delay variability, with the passgate-based families displaying especially strong threshold voltage sensitivity. Gate length and supply voltage variation were also significant contributors to variability, with oxide thickness and gate width presenting little effect.

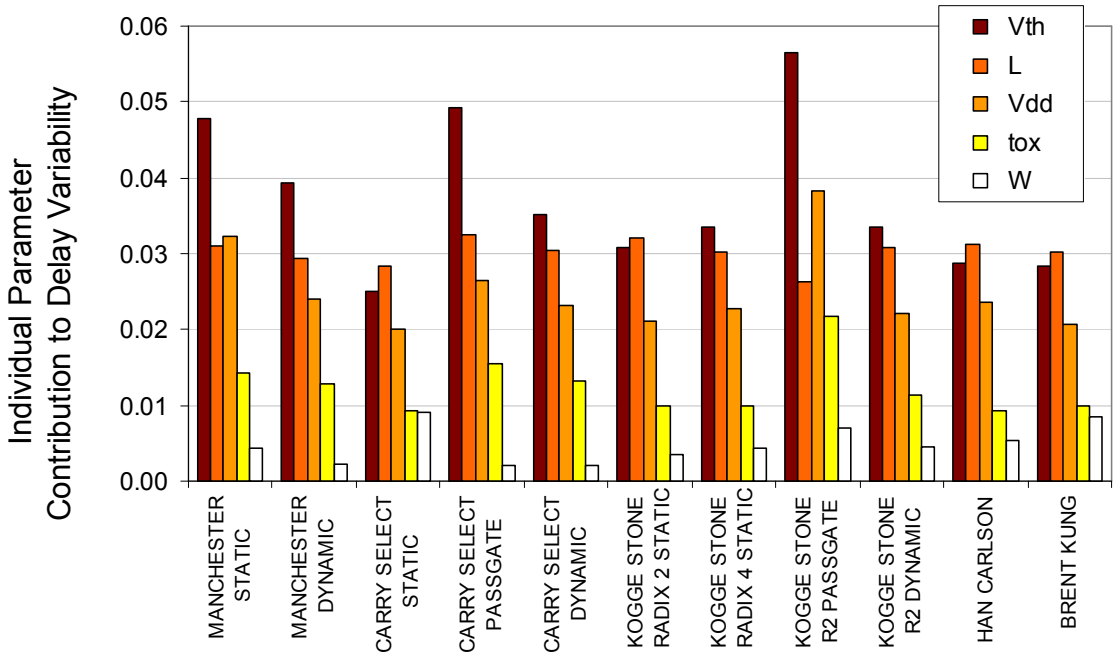

Figure 4.15. Individual parameter contributions to delay variability of 16-bit adders, shown in units of sigma/mean.

As shown in **Fig. 4.16**, variability in active power dissipation was most strongly affected by supply voltage variation:

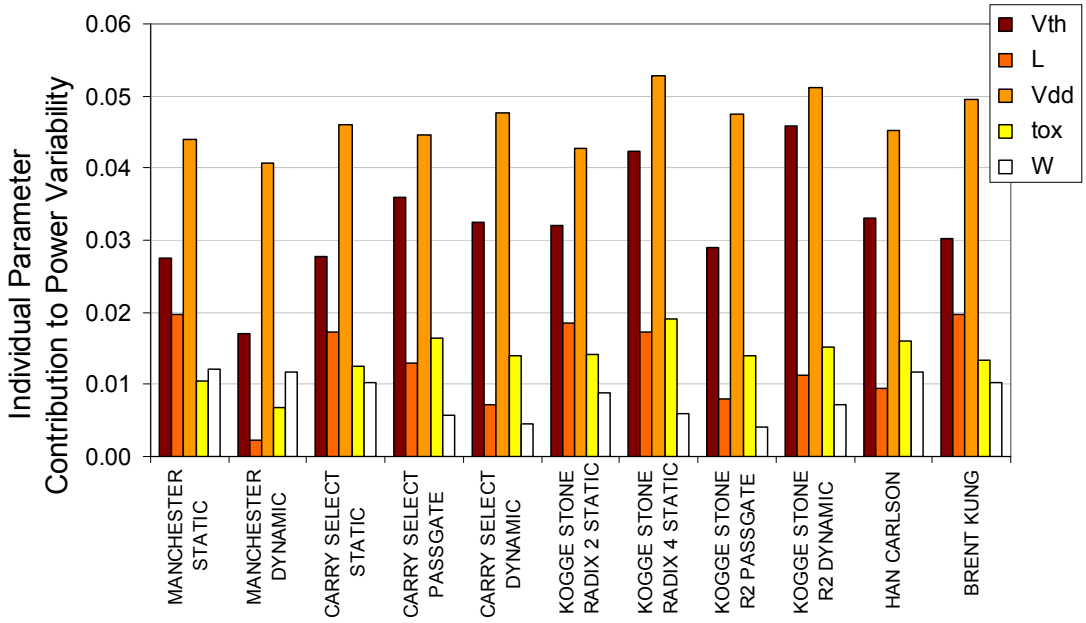

Figure 4.16. Individual parameter contributions to power variability of 16-bit adders, shown in units of sigma/mean.

For the circuit topologies studied in this work, static CMOS circuits were identified as most tolerant of parameter variation, by a narrow margin. Meanwhile, passgate-based circuits suffered substantially larger delay variability than corresponding static implementations, consistent with the observed significant dependence of delay variability on V_{th} variations. The results for power variability of the adder circuits indicated a strong dependence upon intermediate node fanout as designs with larger fluctuating capacitances on internal nodes generally yielded the most widely varying power, while designs with both fewer transistors and more balanced internal signal fanouts displayed the least amount of power fluctuation.

References

[4.1] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling Within-Die Gate Length Spatial Correlation for Process-Design Co-Optimization," *Design and Process Integration for Microelectronic Manufacturing III,* Proceedings of SPIE vol. **5756**, pp.178-188, 2005.

[4.2] R. Wang, P. Friedberg, K. Bernstein, D. Pearson, M. Ketchen, W. Haensch, "De-convolving Variability in Technology/Circuit Co-Design." Research Report# RC23586, IBM TJ Watson Research Center, April 12, 2005.

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

[4.4] Nagel, L. W., and Rohrer, R. A., "Computer Analysis of Nonlinear Circuits, Excluding Radiation", *IEEE Journal of Solid State Circuits*, **SC-6**, pp. 166-182, 1971.

[4.5] BSIM SOI Device Model; http://www-device.eecs.berkeley.edu/~bsimsoi/

[4.6] Y. Cao, et al., "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," IEEE Custom Integrated Circuits Conference, pp. 201-204, 2000.

[4.7] C.-L. Chen and G. S. Ditlow, "Pulse Static CMOS Circuit," U.S. Patent no. 05495188, Feb. 1996.

[4.8] K. Bernstein et al., *High Speed CMOS Design Styles*, 1st ed. Kluwer Academic Publishers, 1998.

[4.9] J. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits*, 2nd ed. Prentice Hall, 2003.

[4.10] P. Kogge and H. Stone, "A parallel algorithm for the efficient solution of a general class of recurrence equations," *IEEE Transactions on Computers*, 1973.

[4.11] T. Han and D. Carlson, "Fast area-efficient VLSI adders," in *8th Annual Symposium on Computer Arithmetic*. Como Italy, March 1982.

[4.12] R. Brent and H. Kung, "A regular layout for parallel adders," *IEEE Transactions on Computers*, 1982.

Chapter 5

Analytical Macromodel-Based Monte Carlo Simulation Framework

In Chapter 3, a thorough characterization of spatial variation in critical dimension was presented. In Chapter 4, traditional SPICE-based Monte Carlo simulation frameworks were introduced and used to perform some initial simulations to investigate the impact of simplified model of CD variation on circuit performance variability. In Chapter 5, the full characterization of CD variation is instantiated in an analytical, macromodel-based Monte Carlo simulation framework to enable a more in-depth exploration of the impact of CD variation on circuit performance variability. After a discussion of the accuracy and execution time required using the updated Monte Carlo simulation framework, two sets of simulations are presented. First, a comparison of various statistical descriptions of the spatial CD variation will is performed. Then, several different process control scenarios are examined to gauge their relative effectiveness under particular styles of design.

5.1 Macromodel-Based Simulation Framework Design

For large simulation sets, the traditional SPICE-based Monte Carlo simulation framework requires a prohibitive amount of simulation time; to combat this problem, the SPICE-based framework was transformed into an analytical, macromodel-based framework using Matlab. The new framework significantly reduces simulation time by replacing the burdensome SPICE simulation step in the original Monte Carlo framework with a fitted polynomial model function (macromodel) that takes as its inputs the properties of each gate in the canonical circuit as well as properties of the incoming signal waveform. Some accuracy is lost by using macromodels in lieu of SPICE, but this inaccuracy is tolerable when the goal is to understand the nature of statistically large quantities of Monte Carlo simulations. The objective for this Monte Carlo framework is to be able to reconstruct reliable distributions and their relevant moments. For example, we can impose some accuracy restrictions on the polynomial function substituted for the SPICE simulation step such that the mean result of n Monte Carlo simulations is no more than ε from the true mean. To do this, we assume the polynomial residuals are normally distributed as $N(0,\sigma)$ and impose a restriction on s such that:

$$\varepsilon \geq k\sigma / \sqrt{n} \tag{5.1}$$

The factor k is a somewhat arbitrary coefficient used to tune the range of ε to a specified number of standard errors. To apply some possible numbers, if we wish to limit ε to 1% of the actual mean, then for n = 100 and k = 3, s would be bound to be no greater than about 3% of the actual mean. In practice, we will see that our σ is actually almost an order of magnitude smaller, at about 0.25% for a single stage; additionally, since we're

stringing several stages together to compose a single canonical critical path, this sample error will be reduced further (by the root of the number of stages) and by the large *n* made possible through quick simulation using macromodels. Thus, we can expect to easily satisfy an ε of far less than 1%.

5.1.1 Model Generation

To begin development of the analytical framework, the analysis of a two-input NAND chain was reproduced. The first step was to develop a reasonably accurate macromodel [5.1, 5.2] for the delay of a single NAND-2 gate given a randomly drawn CD assigned to that gate as well as some information regarding the incoming waveform—namely, the input slew. Through waveform inspection, it was determined that a system of four equations adequately captured the behavior of the NAND-2 gate . The system of equations can be broken into two sets of two equations, with one set corresponding to the situation where the incoming signal is a rising edge, and the other set corresponding to the opposite situation—a falling input signal. Within each of the two equation subsets, the first equation gives the delay as a function of CD (in nanometers) and input slew (in picoseconds, measured from the 10%- to 90%-point of the supply voltage), and the second equation gives the output slew (measured in the same fashion as input slew) as a function of the CD and input slew:

$$Delay_i(CD_i, InputSlew_i) = a_d + \beta_{d1} * CD_i + \beta_{d2} * CD_i^2 + \beta_{d3} * InputSlew_i + e_{di} \qquad (5.2)$$

and

$$OutputSlew_i(CD_i, InputSlew_i) = a_s + \beta_{s1} * CD_i + \beta_{s2} * CD_i^2 + \beta_{s3} * InputSlew_i + e_{si} \qquad (5.3)$$

Of course, since the output slew of one gate is just the input slew of the following gate,

$$InputSlew_{i+1} = OutputSlew_i\left(CD_i, InputSlew_i\right), \tag{5.4}$$

and the entire series of delays and slews in a NAND-2 chain can be reconstructed from a vector of assigned CD's along with the initial input slew.

To generate sample data for the purpose of finding the coefficients of the model equations, 2000 SPICE simulations were performed using completely random (zero correlation) values for the initial incoming slew as well as gate length assignments for both the gate under test and the loading gate (**Fig. 5.1**).

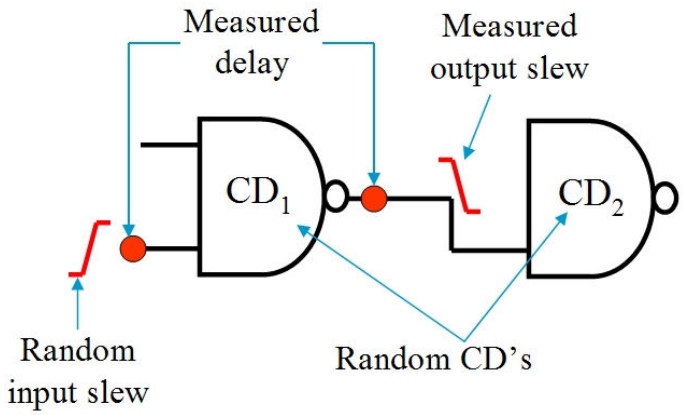

Figure 5.1. Diagram of sample data generation. Random critical dimensions are assigned to each stage, and a random input slew is fed to the designated input to simulate an incoming signal. The delay and output slew are measured in SPICE.

The delay and input/output slews for the gate under test were measured from the resulting waveforms. Although the SPICE results were completely deterministic, the large sample size of the resulting data set made least squares regression a convenient method for estimating the coefficients in the model. In addition, significance testing, though not strictly correct to apply to this deterministic dataset, provided a convenient method for screening for important terms to include in the model. Therefore, ordinary

least squares regression was used to generate estimates of the model coefficients and levels of statistical significance as shown in **Tables 5.1** through **5.4**.

Table 5.1. Model coefficient estimates and levels of statistical significance for the NAND-2 stage delay macromodel created for the case of a rising input waveform.

Delay Model, Rising Waveform

Term	Estimate	Std Err	t Ratio	Prob>|t|
Intercept	-89.779 ps	0.1893	-474.3	0.0000
Input Slew	0.1448	0.0004	347.32	0.0000
CD	3.2615 ps/nm	0.0068	477.44	0.0000
CD^2	-0.0209 ps/nm^2	0.0001	-336.3	0.0000
RSquare	0.999225			

Table 5.2. Model coefficient estimates and levels of statistical significance for the NAND-2 stage output slew macromodel created for the case of a rising input waveform.

Output Slew Model, Rising Waveform

Term	Estimate	Std Err	t Ratio	Prob>|t|
Intercept	-68.299 ps	0.5333	-1.28	0.0000
Input Slew	0.187	0.0012	159.2	0.0000
CD	2.7941 ps/nm	0.0193	145.1	0.0000
CD^2	-0.0141 ps/nm^2	0.0002	-80.09	0.0000
RSquare	0.996323			

Table 5.3. Model coefficient estimates and levels of statistical significance for the NAND-2 stage delay macromodel created for the case of a falling input waveform.

Delay Model, Falling Waverform

Term	Estimate	Std Err	t Ratio	Prob>|t|
Intercept	-11.776 ps	0.1719	-68.49	0.0000
Input Slew	0.159	0.0002	715.69	0.0000
CD	0.9098 ps/nm	0.0063	145.06	0.0000
CD^2	-0.0046 ps/nm^2	0.0001	-80.19	0.0000
RSquare	0.997342			

Table 5.4. Model coefficient estimates and levels of statistical significance for the NAND-2 stage output slew macromodel created for the case of a falling input waveform.

Output Slew Model, Falling Waveform

Term	Estimate	Std Err	t Ratio	Prob>|t|
Intercept	-1.925 ps	0.9642	-2.00	0.0000
Input Slew	0.1787	0.0012	143.39	0.0000
CD	1.2534 ps/nm	0.0352	35.64	0.0000
CD^2	-0.0071 ps/nm^2	0.0003	-21.68	0.0000
RSquare	0.943949			

The R-squared value of each of the four models is very close to one, signifying a good fit.

5.1.2 Accuracy and Efficiency Considerations

Furthermore, by inspecting the residual plots (**Fig. 5.2** and **5.3)**, we can verify that the model is capturing the nonlinear relationship adequately, and at a tolerable level of inaccuracy. From these residual plots, some very slight nonlinear behavior is evident in the residuals for the delay given a rising incoming waveform. However, these residuals have a mean of zero ($8.18*10^{-14}$ps) and standard deviation of 0.094ps; since the mean of the delays in this case is 33.7ps, the nonlinear residual pattern is so small in magnitude that it can safely be ignored. The residual plots for the other three models display no nonlinear behavior at all, indicating that the models are not missing any key higher-order terms that would catch otherwise unaccounted-for systematic effects. For the delay model corresponding to a falling input signal, the residuals also have zero mean ($1.10*10^{-13}$ps) and very small standard deviation of 0.085ps in comparison to the average delay of 32.6ps. In both cases, this single-stage fitting error (roughly 0.25% standard deviation divided by mean) is easily acceptable for purposes of large-sample-size statistical

analysis as explained in Section 5.1. Additional parameters, such as the CD of the loading gate and the device widths, could be included in the model to increase accuracy, but these parameters were found to be less (if at all) significant, and therefore were removed to preserve model simplicity.

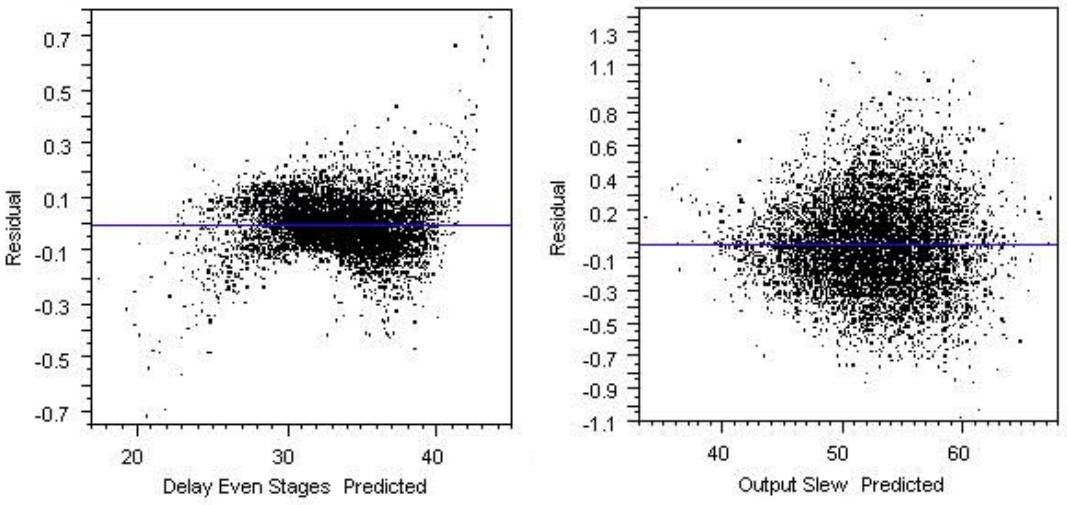

Figure 5.2. Residual plots for the models capturing a single-stage delay and output slew for a rising incoming waveform (all units in picoseconds).

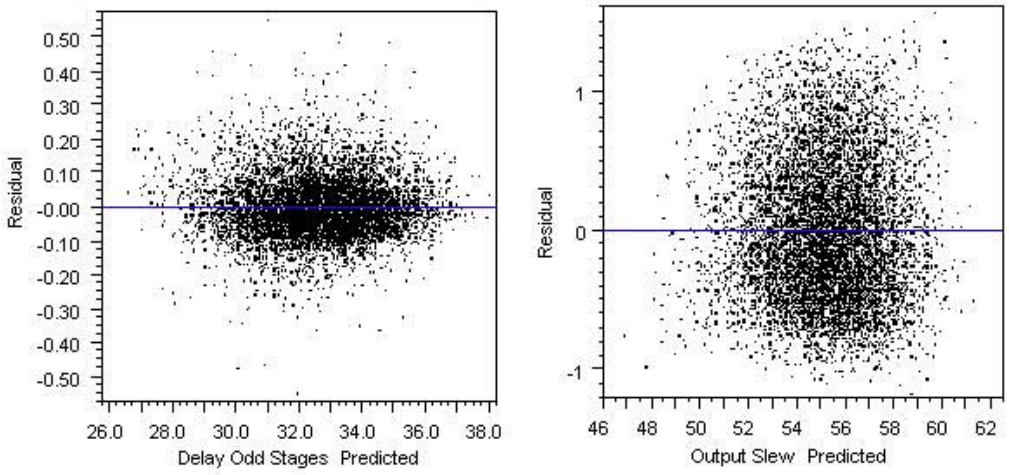

Figure 5.3. Residual plots for the models capturing a single-stage delay and output slew for a falling incoming waveform (all units in picoseconds).

By relaxing the requirements for accuracy slightly in going from a SPICE-based to a macromodel-based framework, significant gains were made in simulation efficiency. A single simulation using the macromodel-based framework requires only 0.25ms providing five orders of magnitude improvement in execution time over the SPICE-based framework. For more complicated canonical circuits composed of several varieties of gates, the model training time will increase since each type of gate will require its own set of macromodels; however, since the training time for a single gate is essentially one set of SPICE-based Monte Carlo simulations, the total training time for any reasonable canonical circuit will be negligible compared to the savings in execution time over the traditional framework.

Since our use for Monte Carlo frameworks is principally to qualitatively explore interplay between manufacturing variation and design performance variability (along with some reasonably accurate quantitative information along with the qualitative), the analytical macromodel-based simulation framework was used for the rest of the work.

5.2 Comparison of Statistical Descriptions of CD Variation

In Chapter 3, it was shown that CD variation could be decomposed into the sum of several systematic variation components and a remaining, marginally auto-correlated random component. However, for our earlier exploration of the impact of spatial variation on circuit performance variability [5.3] presented in Chapter 4, we made a major simplification to this full-decomposition model by assuming that we could capture

all of the spatial information created by the systematic variation components purely through a spatial autocorrelation model based on the original non-stationary CD data. To check the validity of this assumption, we deploy the analytical, macromodel-based simulation framework across several levels of completeness in the statistical characterization of the CD variation.

5.2.1 Candidate Models

Five candidate statistical models of the CD variation were examined, representing each level of sophistication, or completeness, in the decomposition of deterministic components of CD variation:

1) Spatial autocorrelation only, assuming all CD variability is spatially stationary.

$$CD_i = f(\mu, \sigma_{full}, \rho_{full}(\Delta x, \Delta y)) + \varepsilon_i \tag{5.5}$$

2) Across wafer (AW) systematic variation and spatial autocorrelation. The decomposed across wafer variation (**Fig. 3.8**) is simulated separately in addition to the remaining variance (which is described by a mean, variance, and spatial correlation function extracted from the remaining CD fingerprint once the AW component is removed).

$$CD_i = AW_i + f(\mu, \sigma_{full-AW}, \rho_{full-AW}(\Delta x, \Delta y)) + \varepsilon_i \tag{5.6}$$

3) Within-field (WIF) systematic variation and spatial autocorrelation. The decomposed within-field variation component is simulated separately in addition to the remaining variance (which is described by a mean, variance, and

spatial correlation function extracted from the residual CD fingerprint once the *WIF* component is removed). For short critical paths, the placement of the canonical circuit within the systematically varying within-field CD map is randomized.

$$CD_i = WIF_i + f(\mu, \sigma_{full-WIF}, \rho_{full-WIF}(\Delta x, \Delta y)) + \varepsilon_i \quad (5.7)$$

4) Across wafer systematic variation, within-field systematic variation, and spatial autocorrelation. The *AW* and *WIF* components are now simulated separately in addition to the (now much smaller) remaining variance. Again, this residual variance is captured using a mean, variance, and spatial autocorrelation function extracted from the remaining CD fingerprint once the *AW* and *WIF* components are both removed.

$$CD_i = AW_i + WIF_i + f(\mu, \sigma_{full-AW-WIF}, \rho_{full-AW-WIF}(\Delta x, \Delta y)) + \varepsilon_i \quad (5.8)$$

5) Across wafer systematic variation, within-field systematic variation, die-to-die (D2D) systematic variation and spatial autocorrelation. The three systematic variation components are simulated, and the residual variance (depicted in **Fig. 3.10**) is captured by a mean, variance, and spatial autocorrelation function captured by the red (lower) curves in **Fig. 3.11**.

$$CD_i = AW_i + WIF_i + D2D_i + f(\mu, \sigma_{full-AW-WIF-D2D}, \rho_{full-AW-WIF-D2D}(\Delta x, \Delta y)) + \varepsilon_i \quad (5.9)$$

The following section presents the results from the simulations run in parallel using all five candidate models.

5.2.2 Monte Carlo Simulation Results

To enrich the analysis, we simultaneously examine the impact of the path length of the canonical circuit on resulting performance variability by simulating delay variability for path lengths ranging from 0.1mm to 25mm. As is shown in **Fig. 5.4**, the statistical model of CD variation does indeed have a significant impact on the predicted delay variability results. Specifically, as the statistical description grows increasingly accurate by representing additional systematic components as separate from the remaining variance, the variability in delay across the range of path lengths decreases. The bottom curve in **Fig. 5.4**, which captures case (5) from the list above, is the most accurate description of the CD variation and most optimistic prediction in terms of delay variability. The approximation made in Chapter 3 did lead to a relatively sizeable overestimate of levels of delay variability; therefore, complete decomposition of variance is clearly the statistical model of process variation that should be used in future Monte Carlo simulations.

It is also interesting to note that the pathlength at which maximum disagreement between the five candidate models occurs falls roughly in the middle of the range analyzed, at about 10mm. This is not simply a coincidence; rather, it is at this point that the spatial information captured by the within-field deterministic component of variation presents the most significant impact on variability results. At relatively short pathlengths, the stages in the canonical circuit are clustered around one value of critical dimension. At relatively long pathlengths, the stages are strung out across the entire lithographic field, maximizing the averaging affect of CD values and reducing the variability impact. Therefore, in these regimes, capturing the within-field component intelligently only

presents a modest gain in terms of information that can contribute to an accurate prediction of delay variability.

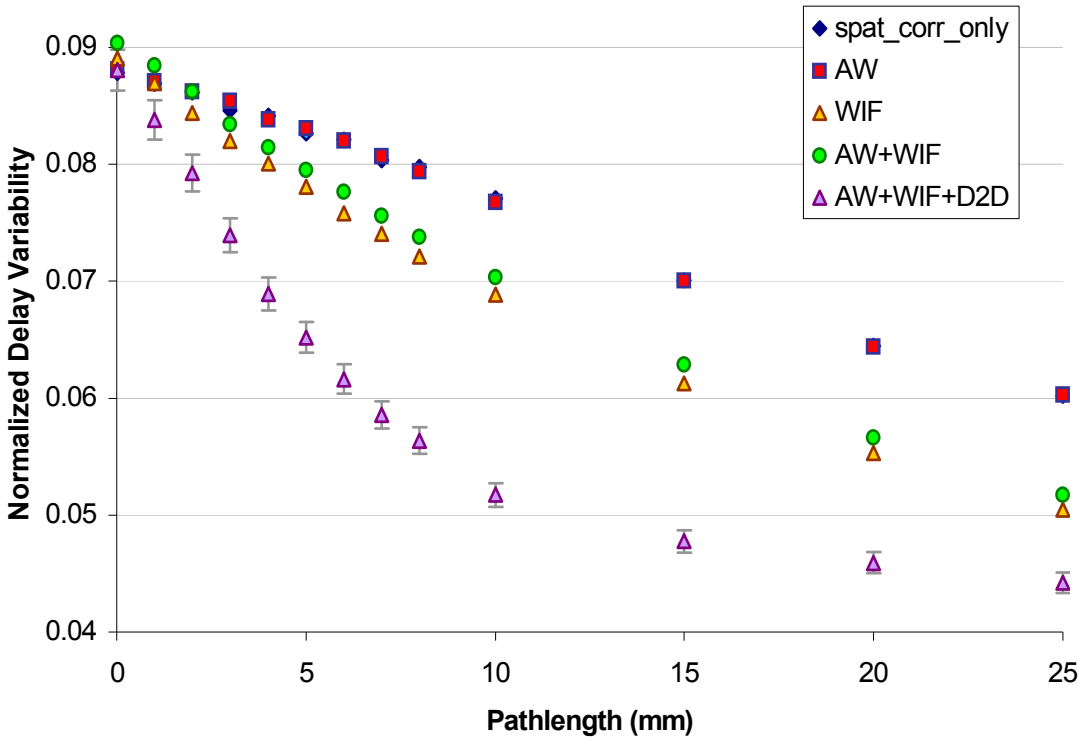

Figure 5.4. Delay variability vs. pathlength for various statistical characterizations of process variation. [5.4]

By contrast, at intermediate pathlengths, the stages are spread over only a portion of the bowl-shaped within-field contour, maximizing the spread in values sampled and therefore the impact of the within-field component of variation. It is at this range of pathlengths that having a model which captures full decomposition of variance is especially valuable. To further illustrate this point, the next section investigates how this macromodel-based simulation framework can be used to evaluate different forms of process control that attempts to partially or fully cancel the various components of variability.

5.3 Monte Carlo for Process Control

Using the statistical model of the process variation that contains a complete decomposition of variance, we can estimate the impact of process control through simulation simply by proportionally reducing the contribution of each variation component in turn. For example, to impose *WIF* process control [5.5], we just reduce the simulated *WIF* component by the desired fraction before adding it to the remaining variation. Through the analytical Monte Carlo simulation framework, process control simulations can be performed very quickly, allowing for an exploration of how the various forms of control (*WIF* or *AW*) impact the resulting circuit delay variability across several candidate critical path lengths. (In this analysis we change the length of the physical critical path while keeping the number of stages, as well as their parasitic loading, constant. This simplification was meant to produce results that are easily comparable.)

As shown in **Fig. 5.5**, the relative impact of the three process control schemes varies depending on the critical path length to a very significant degree. Specifically, for critical paths that span the entire lithographic field, *WIF* process control has a limited impact on circuit performance variability (~6.5% for full *WIF* control), since the bowl-shaped *WIF* variation contour, which contains some natural "averaging", is applied in full to the sequence of path stages. In comparison, the *AW* process control scheme (which we can think of as keeping the entire bowl from moving up and down) is much more effective in combating delay variability for long critical paths, yielding a maximum reduction in delay variability of about 22%.

However, as the path length decreases, the relative impact of *WIF* control grows dramatically. Once the critical path spans roughly one quarter of the lithographic field, *WIF* control becomes the best option in terms of reducing delay variability. At a path length of 1mm, full *WIF* process control leads to a 25% reduction in delay variability in comparison to only ~7% reduction in delay variability from *AW* control. This is also intuitive; as the pathlength begins to sample smaller sections of the bowl, drawn randomly, there is a greater "win" to be garnered by flattening out the bowl than in the case where the critical path spanned the whole bowl.

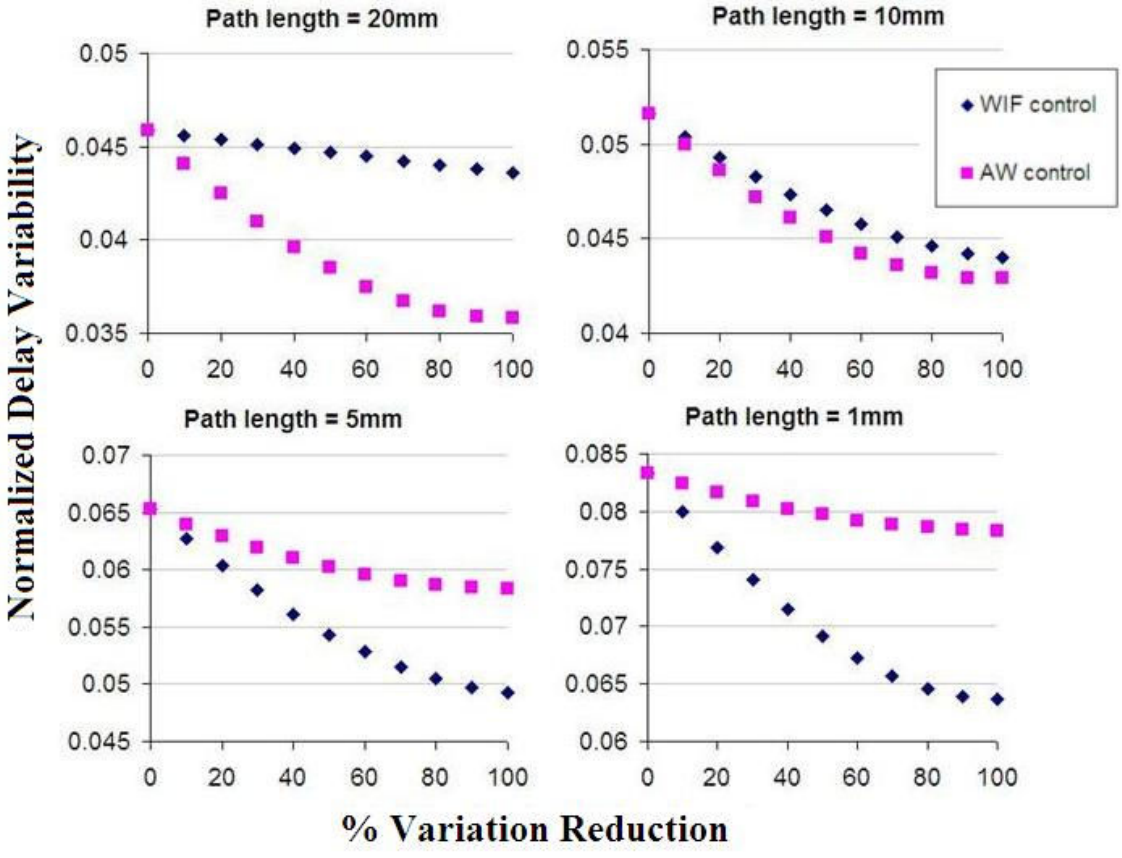

Figure 5.5. Delay variability vs. degree of applied process control, for *WIF* and *AW* control schemes, across a range of critical path lengths (while keeping number of stages and parasitic loading constant).

In sum, these simulations reinforce the usefulness of a statistical description that incorporates full decomposition of variance. And although the palate of Monte Carlo simulations required to sample an illuminating space of the process control arena would require a prohibitive amount of simulation time using a traditional SPICE-based simulation framework, the macromodel-based framework can complete the simulations several orders of magnitude faster. In addition, the formulation of macromodels could easily be expanded to incorporate spatially-based information about variation in other device parameters, such as threshold voltage. The next chapter presents ongoing research into capturing these other significant sources of spatial device parameter variation.

References

[5.1] E. Acar, L. Pileggi, S. Nassif, "A Linear-Centric Simulation Framework for Parametric Fluctuations," *Design, Automation, and Test in Europe Conference and Exhibition*, Proceedings of the IEEE, 2002.

[5.2] A. Kayssi, "Macromodeling C- and RC-loaded CMOS inverters for timing analysis," *6th Great Lakes Symposium on VLSI*, Proceedings of the IEEE, p. 272, 1996.

[5.3] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. J. Spanos, "Modeling Within-Die Gate Length Spatial Correlation for Process-Design Co-Optimization," *Design and Process Integration for Microelectronic Manufacturing III,* Proceedings of SPIE vol. **5756**, pp.178-188, 2005.

[5.4] P. Friedberg, W. Cheung, C. J. Spanos, "Spatial Modeling of Micron-Scale Gate Length Variation," *Data Analysis and Modeling for Patterning Control III*, Proceedings of SPIE vol. **6155**, pp. 61550C1-61550C12, 2006.

[5.5] J. van Schoot, et al, "CD uniformity improvement by active scanner corrections," *Optical Microlithography XV*, Proceedings of SPIE vol. 4691, pp. 304-312, 2002.

Chapter 6

Characterization of Other Sources of Device Parameter Variation

This chapter describes ongoing efforts to characterize additional sources of manufacturing variation. This includes the presentation of the design of new test structures that interleave the novel ELM structures of Chapter 3 with an array of active MOS devices. Anticipated results will be shared.

6.1 Additional Sources of Variation for Study

Now that the spatial model for CD variation is complete, we wish to determine the magnitude of the impact of CD variation on circuit performance variability relative to the impact of other sources of manufacturing variation. Unfortunately, these other sources of variation are substantially more difficult to measure quantitatively because they must be captured indirectly through the electrical performance of active devices. Since there are no direct measurement methods for many of these other device parameters, their effects are convolved with each other and the effect of CD variation itself. To that end, a second set of test structures aimed at decoupling other sources of variation (primarily random

dopant fluctuations and oxide thickness variation) from CD variation has been submitted for manufacture. First, however, the sources of variation other than CD variation will be briefly introduced.

6.1.1 Threshold Voltage

The threshold voltage of a given device is dependent on many device parameters, including dopant distribution, oxide thickness, and gate length. However, in an attempt to treat these device parameters separately, threshold voltage variation will be used in reference to variation in dopant distribution only, a quantity commonly labeled random dopant fluctuation (RDF). RDF is a discrete effect that arises from the random variations in quantity and placement of dopant atoms in the channel region. A common expression for understanding RDF is given by treating the threshold voltage variations due to RDF as independent Gaussian random variables with mean zero, and a standard deviation that depends on the specifics of the manufacturing process, the desired doping profile, and the transistor size, given by

$$\sigma_{Vt} = \left[\frac{qT_{ox}}{\varepsilon_{ox}} \sqrt{\frac{(N_a W_d)}{3L_{\min} W_{\min}}} \right] \times \sqrt{\frac{L_{\min} W_{\min}}{LW}} = \sigma_{Vt0} \times \sqrt{\frac{L_{\min} W_{\min}}{LW}} \tag{6.1}$$

where N_a is the effective channel doping, W_d is the width of the depletion region, T_{ox} is the gate oxide thickness, and L_{min} and W_{min} are the minimum channel length and width respectively [6.1]. From this expression, it is clear that the variation in threshold voltage relative to the nominal threshold voltage value will increase as the size of the device (given by LW) scales to smaller dimensions. At the same time, device threshold voltages decrease with scaling, increasing the relative magnitude of RDF. At the 250nm gate

length technology node, nominal threshold voltage was roughly 0.45V, and threshold voltage variation was 21mV (one-sigma); scaling the gate length to 50nm leads to a nominal V_{th} of 0.2V and σ_{Vt} of 32mV [6.2]. Therefore, the relative magnitude of RDF has more than doubled over that period, highlighting the need for thorough characterization of this growing problem.

Note also that in the treatment of RDF given in **Eq. 6.1**, oxide thickness variation and RDF are intertwined, since the oxide thickness appears in the σ_{Vt0} term; similarly, since the gate length appears in the expression, gate length variation will be convolved to a degree with RDF. The interplay between these device parameters makes it very difficult to separate them spatially; fortunately, as will be shown in the next section, oxide thickness variation for the technology generation being targeted in this work will have negligible impact on RDF. Therefore, oxide thickness variation and RDF will essentially be lumped together, and an attempt will be made to characterize gate length independently so its variation can be separated from the variation of other device parameters.

6.1.2 Oxide Thickness

Oxide thickness variations are generally low spatial frequency, across-wafer and wafer-to-wafer variations that are induced by drift in thermal gradients during the oxidation process and mismatch between rapid thermal oxidation tools. At the scale of current technology, the magnitude of oxide thickness variation is commonly limited to well less than a 10% 3-σ level across the entire wafer [6.3], even with a typical gate oxide thicknesses of only a few nanometers. Since this level of variation is much smaller than

the variation in threshold voltage and gate length, this source of variation is often simply convolved with the other sources, as it will be with RDF in this work.

As device sizes scale to the end of the silicon roadmap, however, oxide thickness variations will begin to exhibit a random and significant intrinsic fluctuation when the correlation length of the Si/SiO_2 interface roughness becomes comparable to the characteristic dimensions of the device [6.4]. The effect on threshold voltage will be similar to that of RDF in that the effect will be entirely random and will be comparable in size to RDF once the gate length drops below roughly 30nm. By the time those device sizes are in production, though, high-*k* dielectrics will have replaced SiO_2 to enable thicker gate dielectric layers and therefore lower relative dielectric thickness variation due to surface effects. Additionally, our experimental data come mostly from a 90nm process; at this technology node, interface roughness has a negligible impact on local oxide thickness variation, and we can safely assume that lumping oxide thickness variations and random dopant fluctuations together is a reasonable approximation.

6.1.3 Interconnect

Variations in interconnect resistance and capacitance are a rapidly growing problem in VLSI design. A dramatic increase in difficulty of creating a single back-end layer is the major culprit; instead of a single patterned metal layer surrounded by a simple oxide dielectric, modern back-end design calls for low-resistivity copper metal to be used in conjunction with a low-*k* dielectric, with a barrier metal used as an interface between the conductor and insulator. Since copper is difficult to etch, dual-damascene processing has

become the standard, where chemical mechanical polishing is required to grind a blanket layer of copper back into shapes defined by an underlying, pre-existing dielectric mold. Finding a suitable low-*k* dielectric is a significant hurdle as well, since the material must not only have a low dielectric constant but also acceptable structural properties, and film density and Young's modulus decrease linearly with the dielectric constant *k*. Recent advances include a chemical vapor deposition precursor for the dielectric, which adds structural integrity without compromising the dielectric constant, and the introduction of high-concentration pores that reduce the dielectric constant further [6.5]. Finally, just as with the rest of the chip, interconnect dimensions are being aggressively scaled, which serves to increase the proportion of total interconnect material that is composed by the barrier metal (which does not scale) [6.6]. The barrier metals generally have a much higher resistance than copper, so variations on barrier metal thickness have a strong impact on interconnect resistance. Although TaN barrier metal films can be deposited to 1nm uniformity using atomic layer deposition, a small variation in linewidth has a magnified impact on resistance variation.

6.2 Decoupling CD Variation from Other Sources of Variation

This section describes a second set of test structures designed for foundry manufacture that target the characterization of spatial variation in sources of variation other than CD variation. Since other sources of variation are convolved with critical dimension in the electrical performance of devices (arguably the best way to collect data

on other sources of variation), these test structures must be thoughtfully designed to allow for decoupling of CD variation from the other sources.

6.2.1 Novel Active-Device Test Structures

This second set of test structures principally consists of a 16-by-300 array of MOS transistors adapted from the work of Shimizu et al [6.7], as shown in **Fig. 6.1**. The array is composed of the MOS transistor cells themselves, on which the rigorous spatial measurements will be performed, as well as a surrounding frame of circuitry designed to allow the user to access a single transistor at a time. In the access circuitry, pulses are sequentially propagated through an individually-driven row and column of clocked elements. When the pulses arrive at a particular stage of the row and column of clocked elements, a number of functions are enabled for the individual MOS cell within the array that is indexed by the selected row and column. First, a multiplexer along the row of clocked elements is toggled to allow full control of the gate of an individual MOS cell through the pad marked "Gate" in **Fig. 6.1**. Second, a transmission gate along the vertical axis is switched to allow a current path through the device, from the node marked "Source force" to the node marked "Drain force." To reduce the number of pads in the design, the source force pad is tied to V_{dd} and the drain force pad is tied to ground. Finally, the "Drain sense" and "Source sense" nodes are enabled through multiplexers so that the drain voltage and source voltage of the targeted device can be measured in turn in four-point Kelvin fashion.

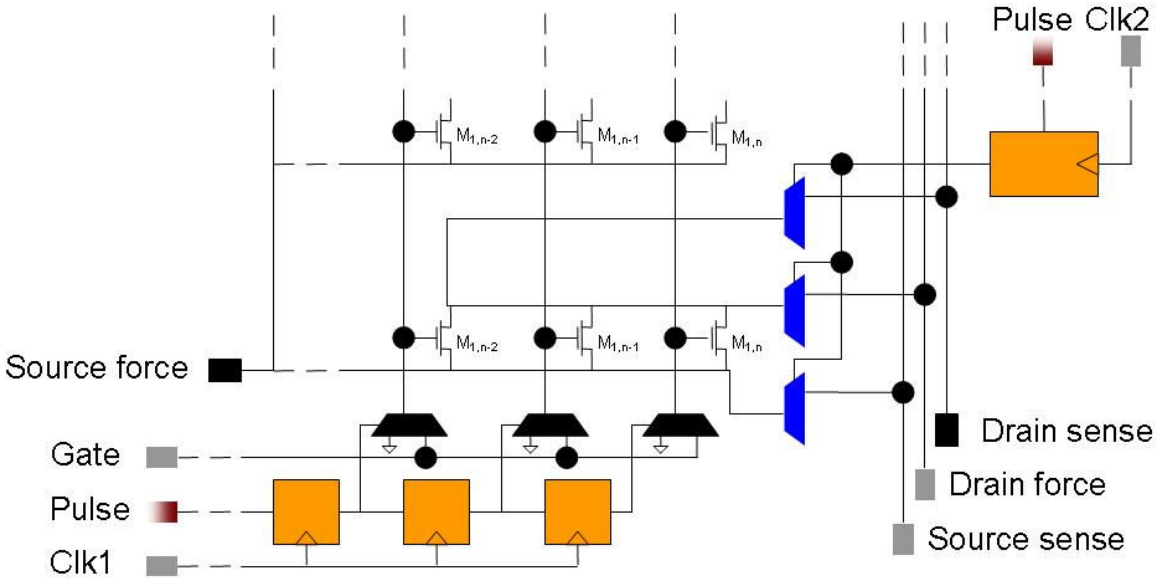

Figure 6.1. Diagram of the general MOS array design, adapted from [6.7]. MOS devices within the array are accessed in one-at-a-time fashion using a series of pulses controlled by clocking elements in peripheral logic.

In practice, the source voltage will always be equal to the supply voltage (V_{dd}), but the drain voltage will be greater than 0V since there is a voltage drop across the transmission gate along the current path. Thus, by measuring both the current through the device at various gate voltages as well as the drain voltage of the device, a standardized threshold voltage can be captured. This procedure will be visited in more detail later.

The Cadence layout of a single MOS cell from the bottom right-hand corner of the array is shown in **Fig. 6.2**, along with the associated transmission gates and multiplexer. Within the MOS cell, two dummy gates are placed on each side of the device to provide uniform lithographic pattern density throughout the array. The multiplexer at the bottom of the figure allows for access to the gate of the device, and the transmission gates at the right-hand side of the figure allow current to flow through the device and for the source voltage and drain voltage to be independently measured.

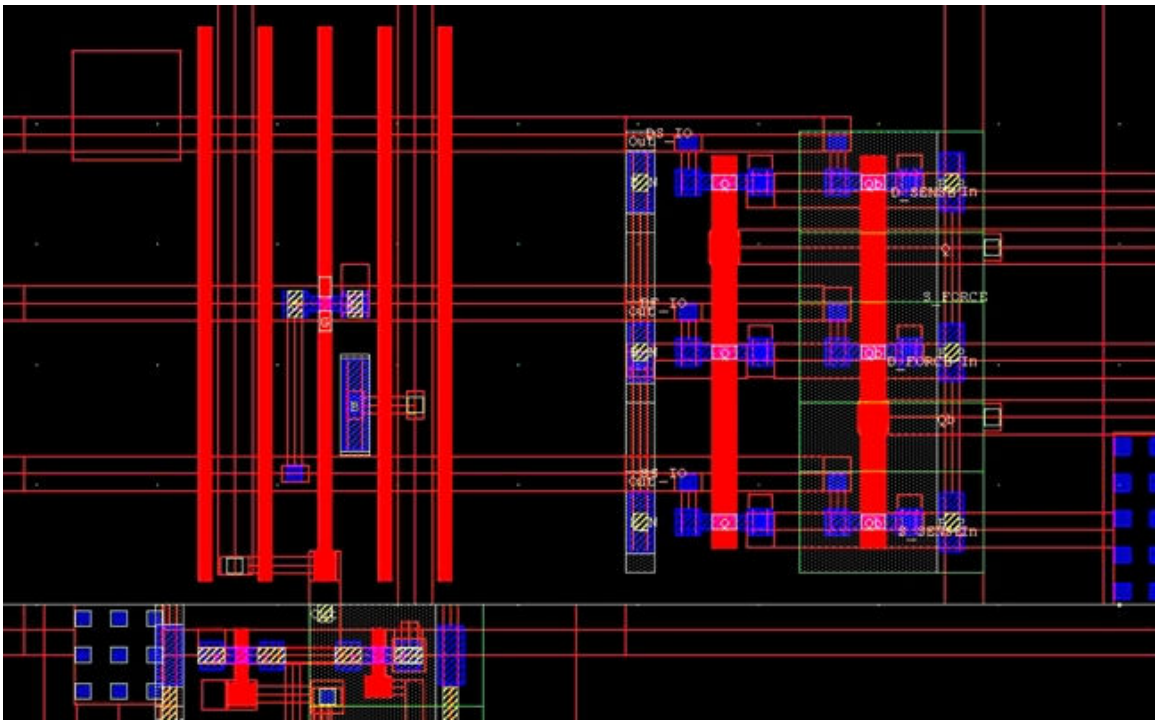

Figure 6.2. Cadence snapshot of a MOS cell and associated access logic. The structures on the right hand side are the three transmission gates connecting a row of the array to the drain sense, source sense, and drain force nodes. The structures at the bottom left comprise the multiplexer through which the gate node access the polysilicon gates of devices in the selected column of the array. The series of 5 vertically-oriented polysilicon structures are the gate of the target device (center line) and two pairs of dummy gates to provide pattern density uniformity.

Several design choices were made to increase the accuracy of measurements on the MOS array cells. The primary potential source of error lies in allowing alternate paths for rail-to-rail current through non-targeted devices via leakage. The multiplexer along the clock row allows the gate node to control the voltage of all gates within the selected column. Therefore, all these devices may exhibit some leakage since their gates are active, despite having zero source-to-drain bias. Additionally, the rail-to-rail source-to-drain bias set up across the targeted device also exists across the other 299 devices within the array row selected by the clock column; therefore, these devices may exhibit some leakage despite having no voltage applied to the gate. Through simulation it was

determined that the former case was of greatest concern—with the voltage at the gate node being applied to the 15 devices in the selected column intended to be inactive, the accuracy of the intended measurement relies on the ability of the access transmission gates on intended off-current paths to have very low leakage. If the access transmission gate is designed to have large drive power when turned on but low leakage when turned off, the problem has been mitigated. To limit the current along unintended paths, the sizes of transistors making up the access transmission gates were skewed so that the gate length was increased to 200nm (double the nominal condition) while keeping the transistor width at the nominal (120nm) value. Since leakage varies exponentially with gate voltage, this sizing strongly reduces leakage current without compromising drive current. Simulations indicate that the total leakage current is on the order of 10,000 times smaller than the current in the targeted device, meaning that variation in other, non-targeted devices in the array will not significantly confound measurements on the targeted device.

In addition to the MOS electrical structures, a means for collecting independent CD variation information is required to allow for the CD variation to be accounted for separately and decoupled from threshold voltage variation and other variation sources. To accomplish this task, the serpentine ELM structures from Chapter 3 are woven into the MOS array at regular intervals. By interleaving the ELM structures with the MOS devices and then measuring them side-by-side, the CD variation from the other sources of variation contributing to overall device performance variability. An annotated, global-view snapshot of the Cadence layout for the MOS array with interwoven ELM structures is shown in **Fig. 6.3**.

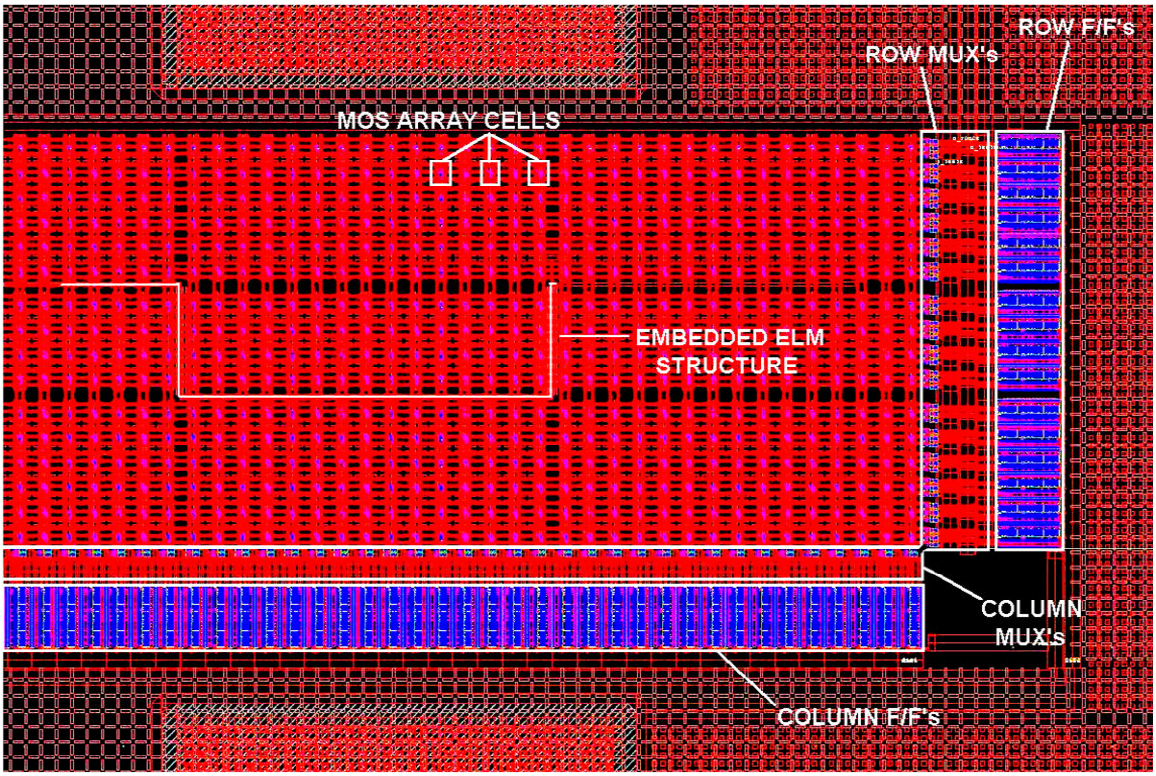

Figure 6.3. Global-view snapshot of MOS array layout with ELM structure embedded.

A zoomed-in snapshot of a single measurable section of the embedded ELM structure surrounded by several cells of the MOS array is shown in **Fig. 6.4**. Each line under test within the ELM structure is surrounded by two dummy lines on each side to match the lithographic pattern density of the gate of each MOS cell; this way, a fair correspondence may be established between variation in the linewidth of the ELM line under test and variation in the CD of the MOS cell gates.

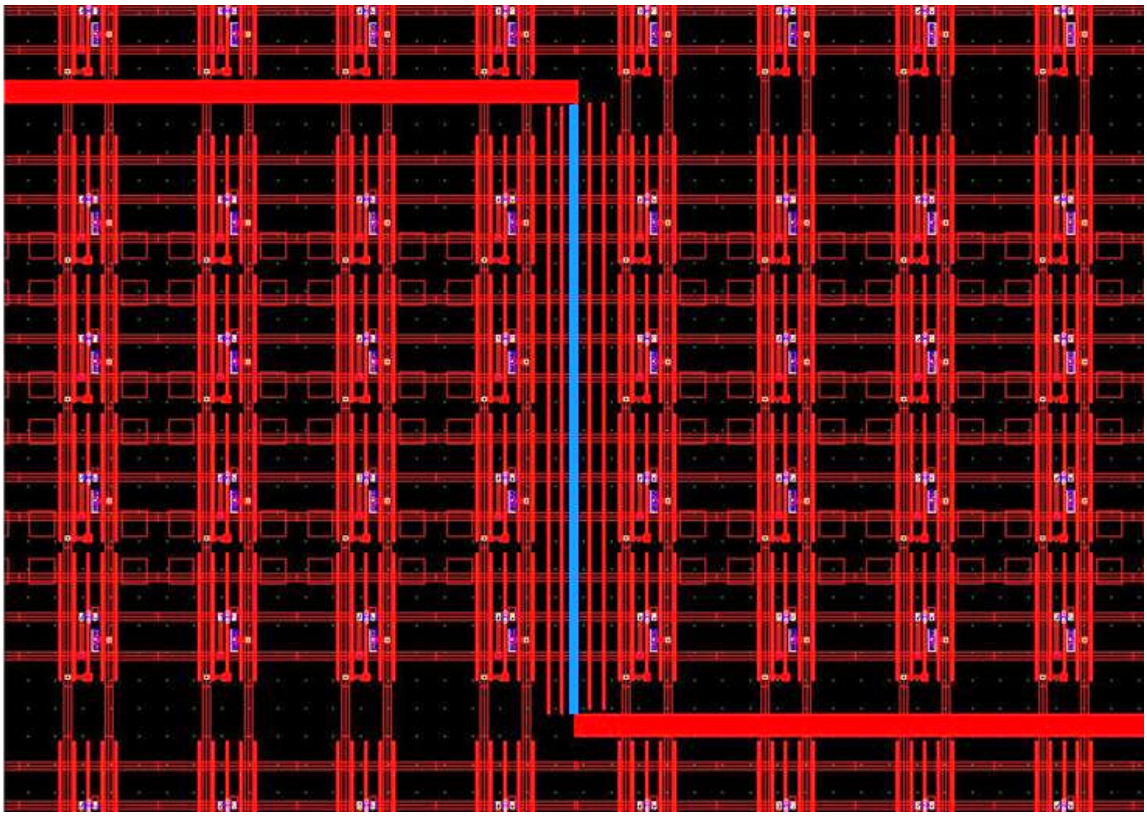

Figure 6.4. Close-up of ELM structure embedded in MOS array. Contacts are made to the two thick horizontal sections and current is passed through the 22µm-long vertical section for the CD measurement. Individual cells of the MOS array surround the ELM structure.

In addition, this test structure design contains several standalone MOS devices that are identical to those in the array except that they are equipped with four dedicated pads to gate, source, drain, and bulk to allow for independent characterization of a MOS cell. Similarly, individual instances of a single line under test identical to those in the ELM test structure are incorporated into the design to allow for independent characterization of the ELM test structure. Finally, Van der Pauw structures are scattered throughout the design to allow for measurement of sheet resistance that can be used in conjunction with ELM test structure results to back out a direct measure of linewidth as described in Chapter 3, Section 3.1. A schematic of the entire chip layout is shown in **Fig. 6.5**, with

the multiple instances of the MOS array with embedded serpentine ELM structure depicted by long rectangles inscribed with the designed-for nominal gate length (100, 120, or 140nm):

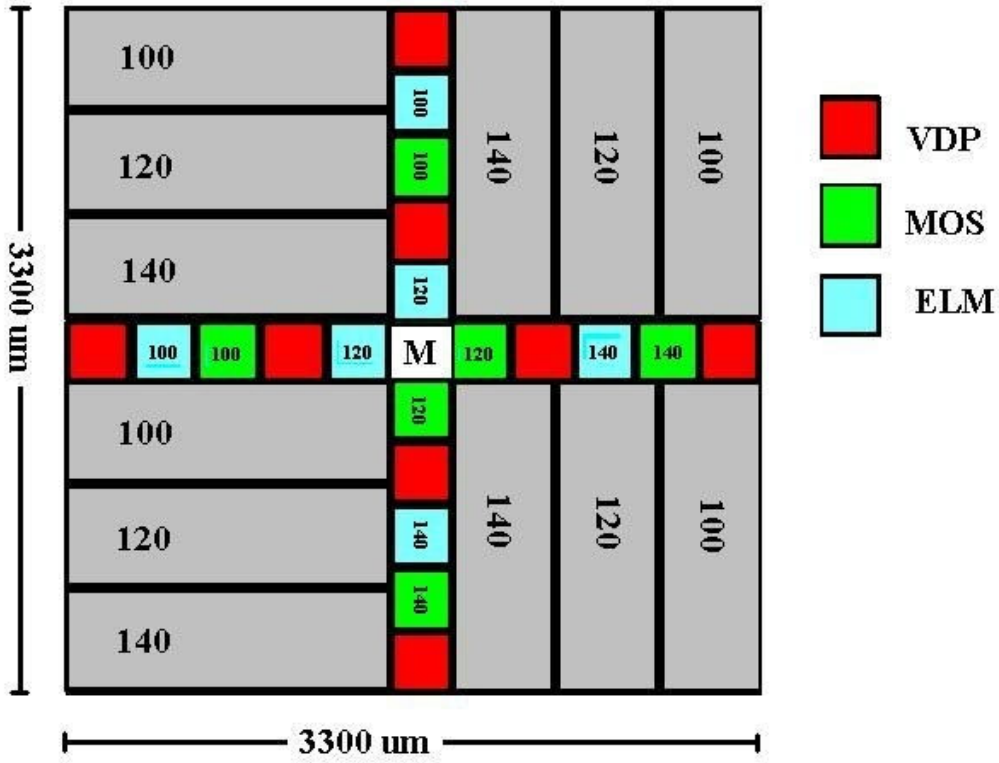

Figure 6.5. Schematic of the second test structure chip layout. The long rectangles denote instances of the MOS array embedded with the serpentine CD-only ELM structure. Each instance is inscribed with the nominal designed-for CD of all devices within the given instance. The instances of the stand-alone Van der Pauw, MOS cell, and ELM structures are also denoted.

6.2.2 MOS Array Characterization Method

As described in Section 6.2.1, several design choices were made to control the level of leakage current through non-targeted devices. The resulting sensitivity to variation in gate length of the targeted device should be unaffected by variation in non-targeted

devices within the array. However, by keeping the width of the access transmission gates at nominal values, a slight oversight was made with regard to management of parasitic

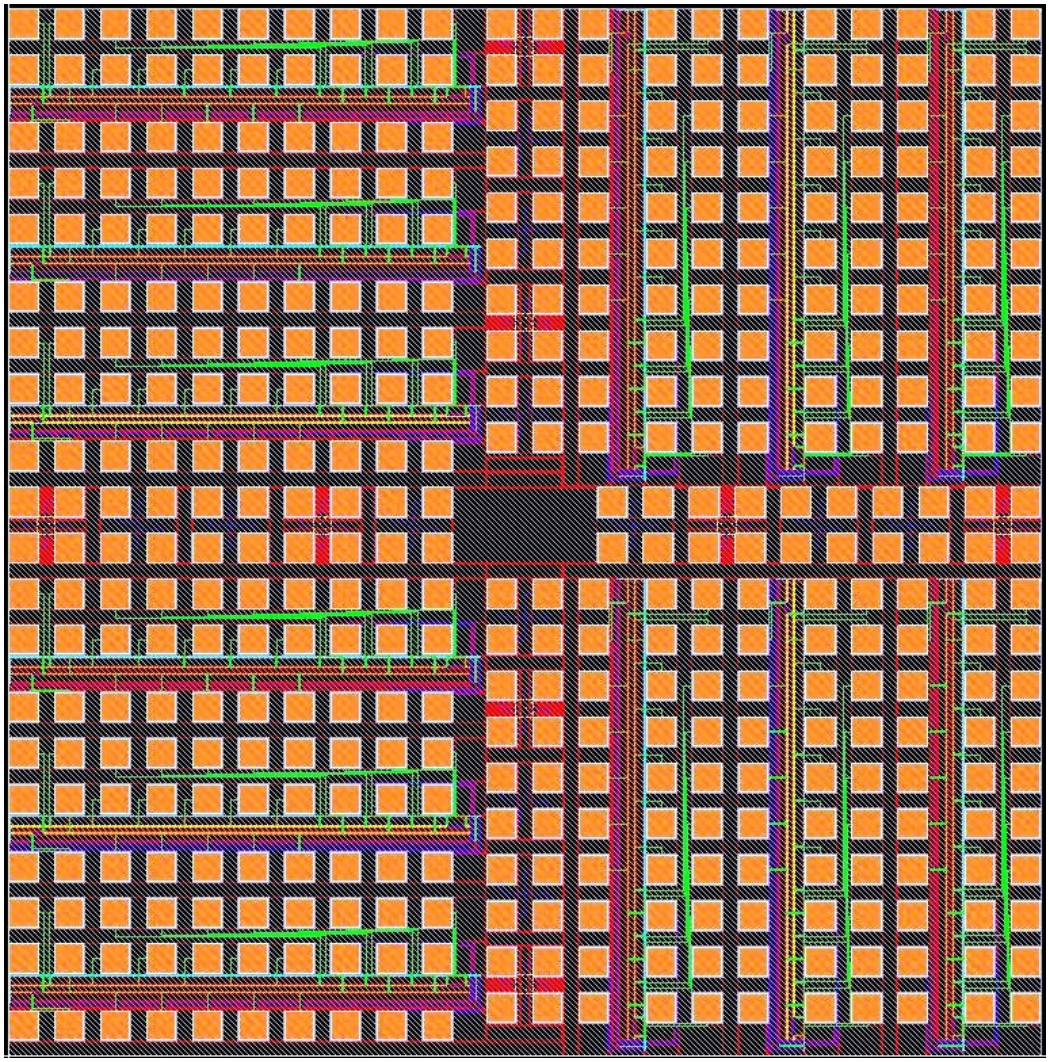

Figure 6.6. Full-chip Cadence layout of the second test structure chip. The 3-by-10 pad frame of each instance of the MOS array embedded with serpentine ELM structure is repeated 12 times, 6 times for horizontal instances of varying nominal CD, and 6 times for vertical instances of varying nominal CD. The 2-by-10 pad frame of the standalone devices, Van der Pauw structures, and ELM structures is repeated 4 times, twice each in vertical and horizontal orientations.

resistances along the current path to the intended device. That is, by keeping the transistors within the access transmission gate only as wide as the targeted device, the resistance of the transmission gate is comparable to that of the targeted device. Because

of this significant resistance, there is a substantial voltage drop across the transmission gate since it lies in the current path. Consequently, the drain voltage "seen" by the targeted MOS array device will be higher than ground. To illustrate this effect, the drain voltage and device current are plotted against the applied gate voltage in **Fig. 6.7**.

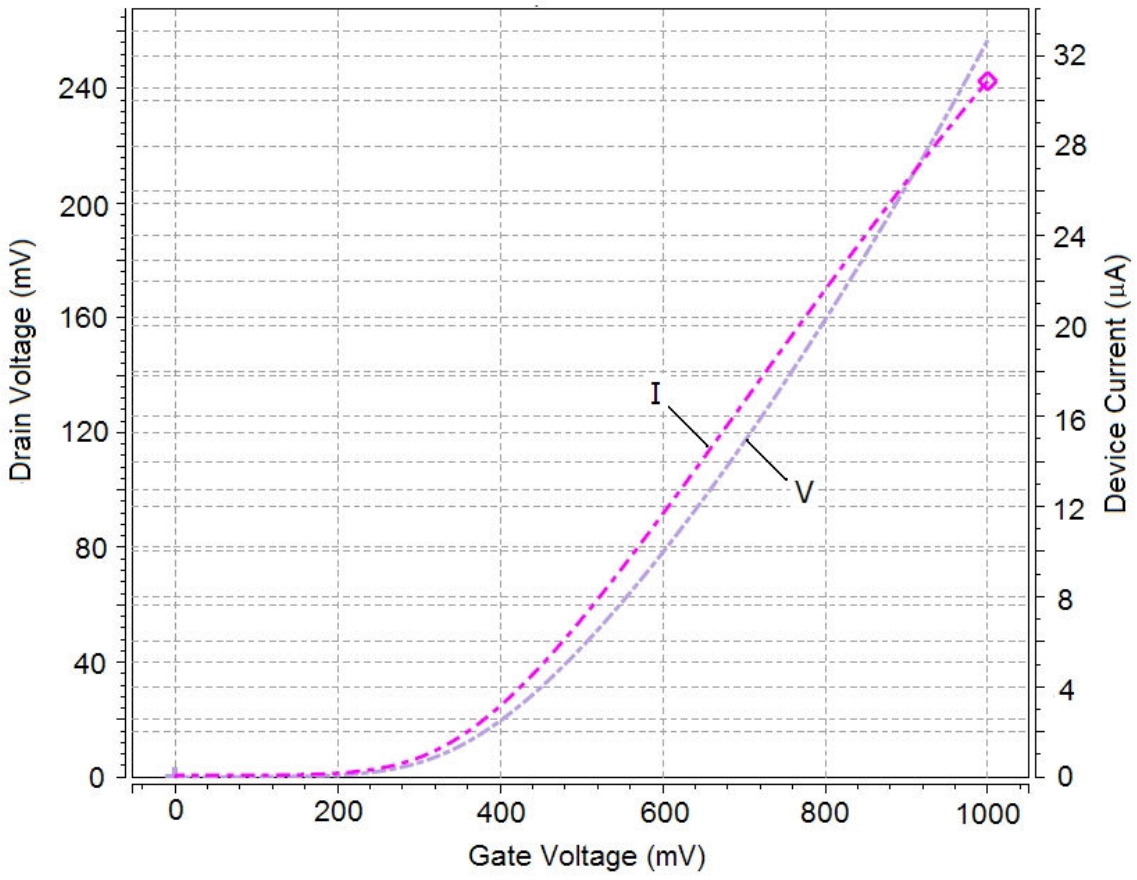

Figure 6.7. Plot of drain voltage and device current against the applied gate voltage. As the device turns on, the device current increases and the voltage drop across the transmission gate in series with the device increases. As an upshot of this voltage drop, the drain resistance rises from the desired value of 0V.

Furthermore, the degree to which the drain voltage rises and causes body effect will depend on variations in the voltage drop across the transmission gate, which in turn will depend on variations in the devices that compose the transmission gate. As shown in **Fig. 6.8**, corner simulations of ±10% variation in the gate length and width of the devices

within the access transmission gate indicate that the current measured will depend strongly on the device properties of that particular access transmission gate. The total spread in corner simulations represents about 10% of the measured current.

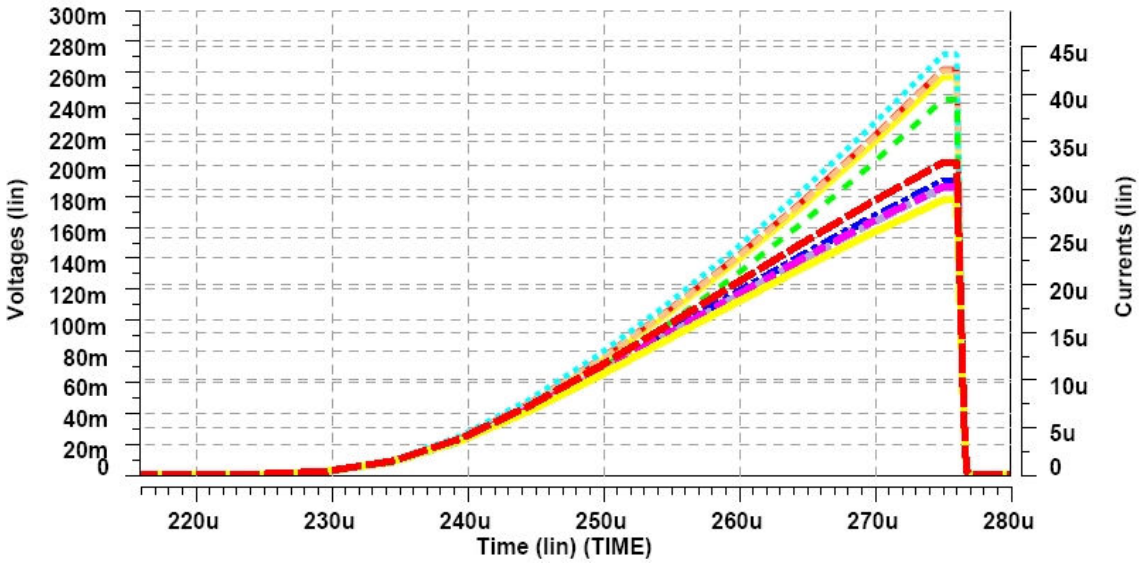

Figure 6.8. Plot of sense voltages (upper curves, left hand vertical axis) and device current (lower curves, right hand axis) through a sweep of the gate voltage from ground to V_{dd}, with +/-10% variation applied to the gate length and width of the devices in the access transmission gate.

Fortunately, since the drain voltage that is seen by the targeted device can be directly measured through the drain sense node, we can still avoid confounding a measurement on the targeted device with variations in the access transmission gate. By measuring the current, whatever it happens to be, along with the drain voltage of the targeted device, simple division will provide the resistance of the targeted device. First, we can plot the voltage drop across the transistor only (by taking the difference between the source voltage and drain voltage, both of which can be measured independently in 4-point Kelvin fashion) as well as the device current against the gate voltage, as shown in **Fig.**

6.9. Then, the resistance can be calculated by dividing the source-to-drain resistance by the device current, as shown in **Fig. 6.10**.

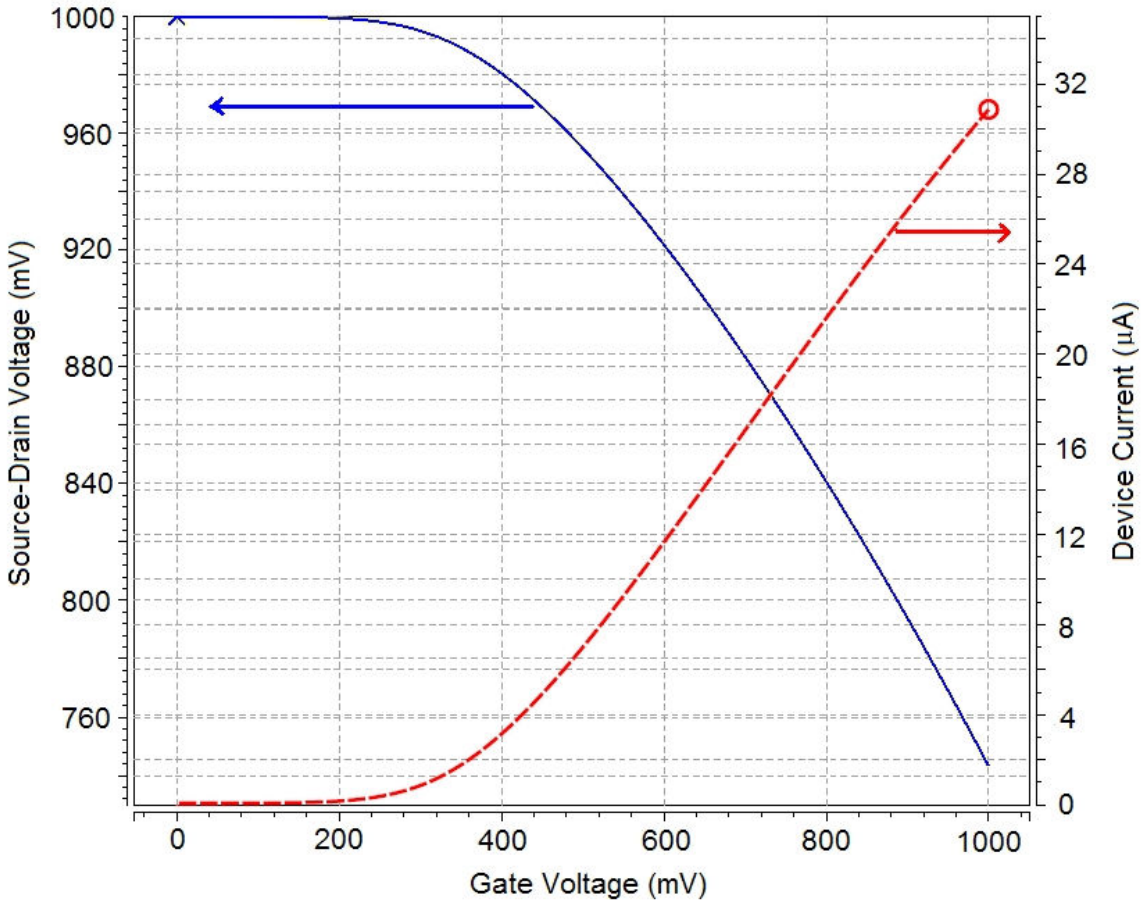

Figure 6.9. Plot of source-to-drain voltage and device current against the applied gate voltage. Since the source voltage and drain voltage are measured independently via 4-point Kelvin method, these curves are independent of the resistance presented by the access transmission gate.

This resistance can be measured for several applied gate voltages to map out a portion of the device resistance vs. applied gate voltage, which in turn can be used to perform a comparison of relative threshold voltages within the devices. However, due to limitations in measurement time (that will be described in the next section), only one device resistance measurement will be taken per targeted MOS device. By selecting a single

gate voltage to apply that is near the expected threshold voltage of roughly 250mV, a "threshold device resistance" figure of merit can be used to compare devices and determine the spatial variability statistics.

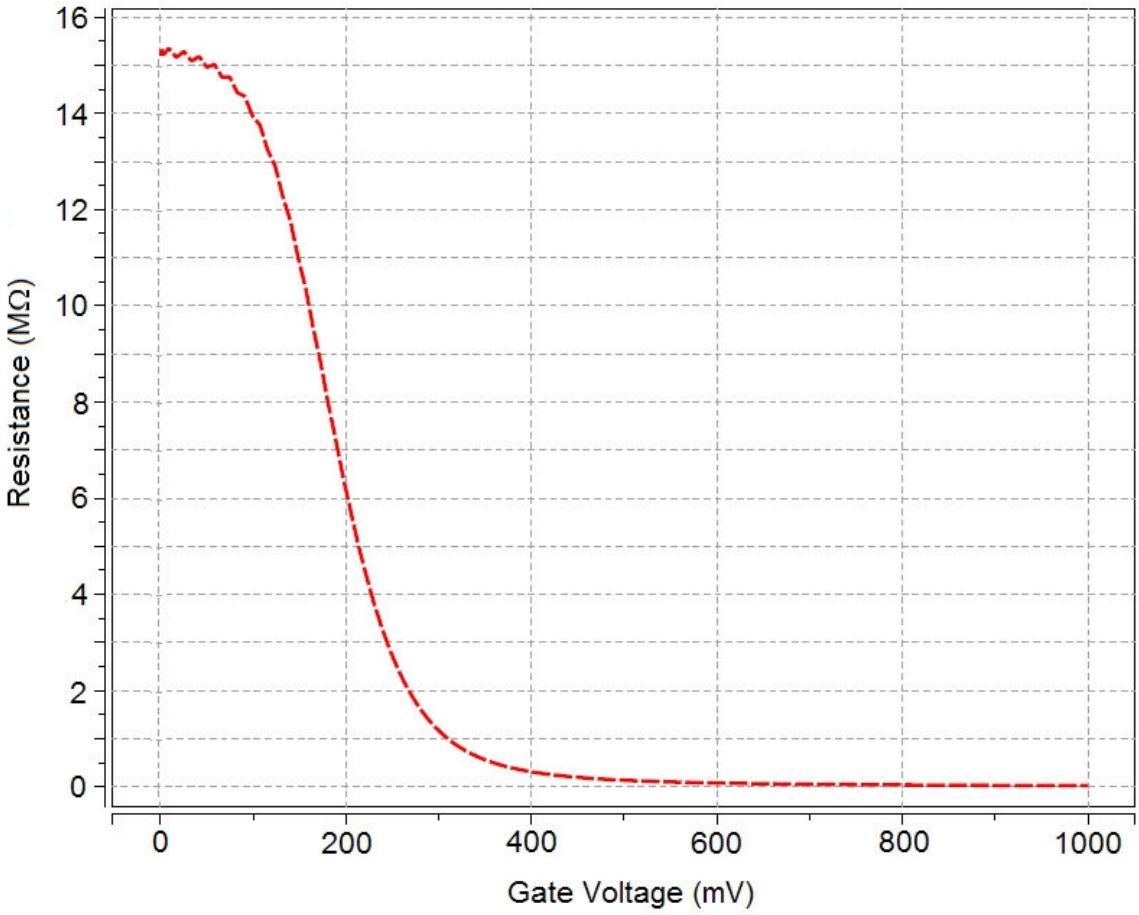

Figure 6.10. Plot of device resistance against the applied gate voltage. The device resistance can be backed out from voltage and current measurements and is independent of all other elements in the MOS array.

6.2.3 Measurement Sampling Plans

In practice, characterizing the MOS arrays would require significant measurement time. With 4800 devices in each array, two instantiations of the array at nominal designed gate length per orientation, and two orientations (horizontal and vertical) per

chip, a total of 19,200 devices are available for dense spatial analysis on each of the 25 chips. By running the measurement program with a "dummy chip" in place of the actual chip, it was determined that a single measurement—which consists of making the appropriate automated hardware connections, generating the intended voltages at various nodes, and measuring the desired voltages and currents—requires roughly 3 seconds. Therefore, the total characterization time for a single chip using the minimum number of measurements (one) per targeted device is roughly 16 hours. Obviously, a reduced sampling plan will be required for a substantial fraction of the measurement workload.

Three measurement plans were created for characterizing the MOS arrays. In the first measurement plan, each device is to be measured a single time, requiring 8 hours for each chip at each orientation (the hardware requires the user to manually rotate the chip 90 degrees between the two orientations). The second measurement plan reduces the number of devices measured by a factor of 2 in each axis of the array. By measuring only every other row and column of the array, the dimensions will effectively be reduced to 8 rows by 150 columns with the separation distance between devices doubled in each direction as compared to the exhaustive measurement plan. Using this plan, with only 1200 targeted devices per effective array and, therefore, 2400 targeted devices per orientation, the measurement time would be only about 3 hours. At this rate, both orientations of a single chip could be measured within a day at the lab. The third measurement plan would be a weekend job, and could therefore consume up to 60 hours of measurement time. This time could be used to make multiple measurements for each targeted device at various gate voltages surrounding the effective threshold voltage. Using 7 measurements at each targeted device, and exhaustively measuring all 4800

devices per array (and therefore 9600 devices per orientation), the measurement routine would require 56 hours.

The measurement time required to fully characterize the serpentine ELM structure as described in Chapter 3, including multiple measurements at each measurable line to increase measurement precision, is negligible in comparison to the time required for characterizing the devices in the MOS array.

6.3 Measurement Results and Analysis

This section describes the results in two subsections: first, the successful measurements acquired from the ELM test structures and standalone MOS devices are presented. Second, the unsuccessful attempts to characterize the MOS array are described and potential sources of malfunction are discussed.

6.3.1 ELM and Standalone MOS Results

Just as with the similarly-designed micron-scale ELM test structures of Chapter 3, the ELM test structures on the second test structure design submission were measured using an Electroglas Automatic Probe Station. Several measurement iterations were combined to increase the precision of the average measurement, as described fully in Section 3.3.2. Again, these test structures were successfully measured with 100% yield, and again there was an offset in average measured CD between the horizontal (131.9 nm) and the vertical (135.2 nm) orientations. Finally, the variance for the horizontal orientation (7.06 nm^2) was greater than that of the vertical orientation (5.58 nm^2). This is composed of 6.59 nm^2

chip-to-chip variance and 0.47 nm^2 within-chip variance for the horizontal orientation, and 5.01 nm^2 chip-to-chip variance and 0.57 nm^2 within-chip variance for the vertical orientation. These levels of chip-to-chip variation are comparable to those of the first test chip; the within-chip levels of variation are significantly lower (by 40% for horizontal and 15% for vertical) due to an improved uniformity in polysilicon density, which alleviates the previously-seen etch microloading effect.

Similarly, the standalone MOS transistor test structures yielded at 100% and matched the expected MOS characteristics seen in SPICE simulation very closely. Figure **6.11** shows the MOS transistor characteristics simulated in SPICE and the corresponding characteristics measured from a typical standalone transistor. The current from the empirical device falls slightly short of the corresponding current from the simulated device, by about 20%. However, the shape of the characteristic matches very well.

In an attempt to characterize the correlation between CD variation and variability in electrical performance of the MOS device, we can examine scatterplots of the chip-average CD along with the current through the standalone MOS device from that chip and orientation at a given gate voltage near the threshold voltage. As shown in **Fig. 6.12** and **Fig. 6.13**, there is very little correlation between the chip-average CD and corresponding MOS current at a 0.3V gate voltage on the same chip, for both the horizontal and vertical orientations of ELM structure and MOS device. In both cases, the R^2 of a linear fit to the data is extremely weak and not indicative of any statistically significant relationship. In **Fig. 6.14** and **Fig. 6.15**, plots of saturation current for the same devices against average CD for each chip are shown; in this case, the R^2 value of the linear fits is substantially higher, but still very low.

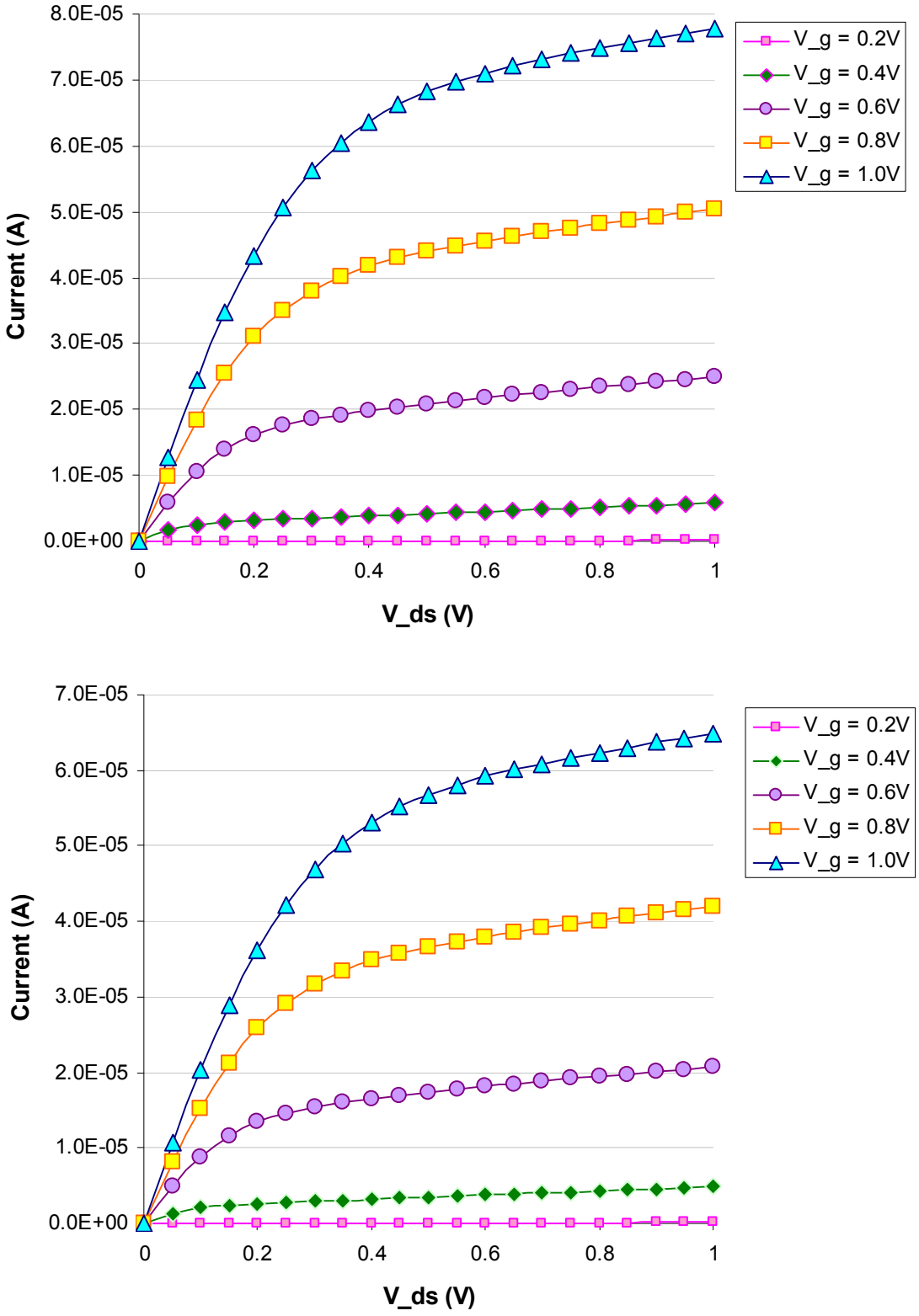

Figure 6.11. Plot of simulated (top) and measured (bottom) single device MOS characteristics.

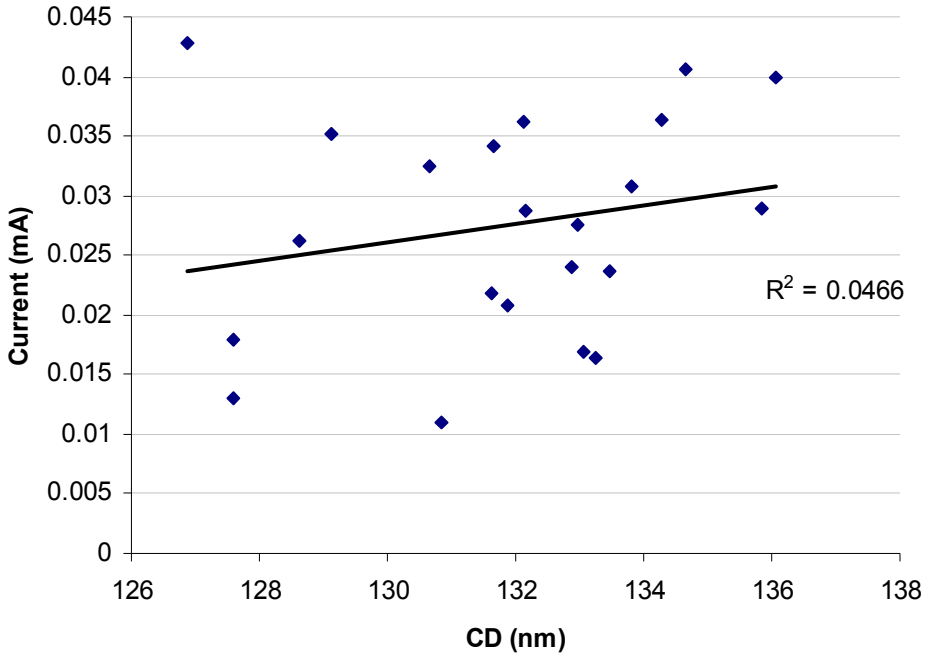

Figure 6.12. Plot of device current through a standalone MOS device (at V_g = 0.3V) vs. average CD for each of the 22 test chips, for the horizontal orientation.

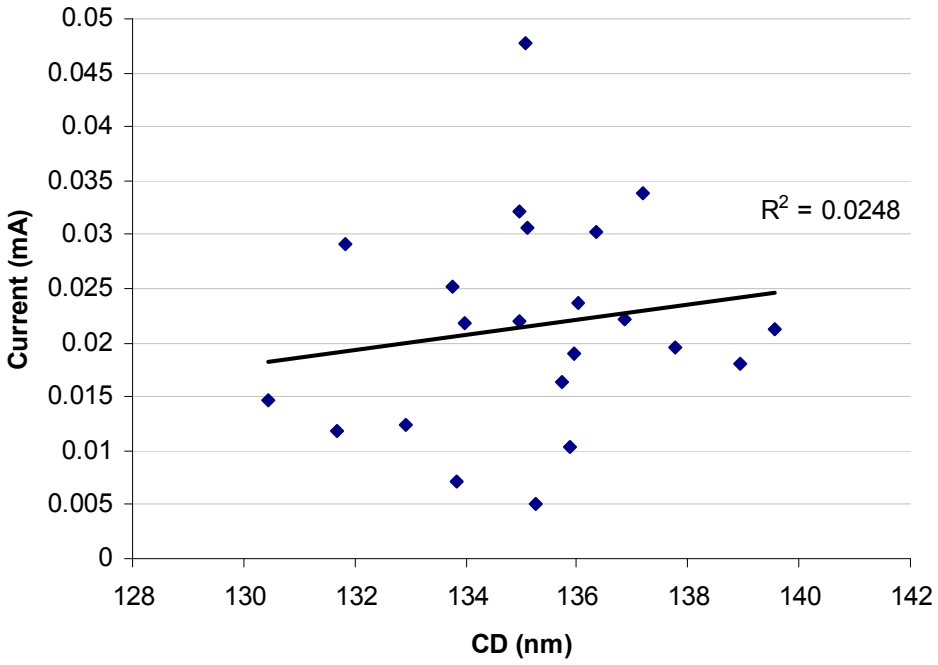

Figure 6.13. Plot of device current through a standalone MOS device (at V_g = 0.3V) vs. average CD for each of the 22 test chips, for the vertical orientation.

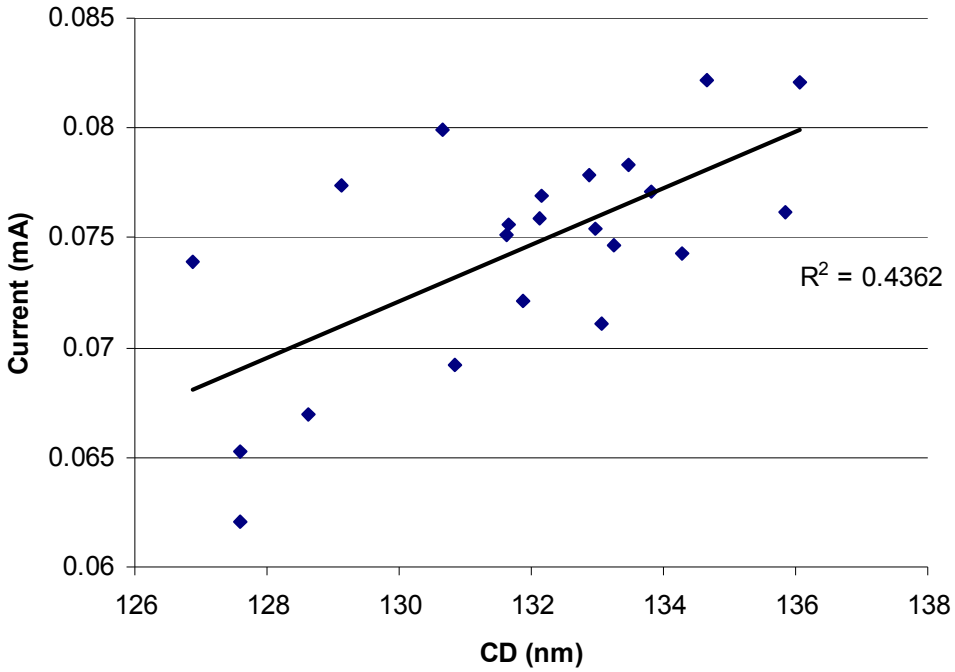

Figure 6.14. Plot of device saturation current through a standalone MOS device (at V_g = 1V) vs. average CD for each of the 22 test chips, for the horizontal orientation.

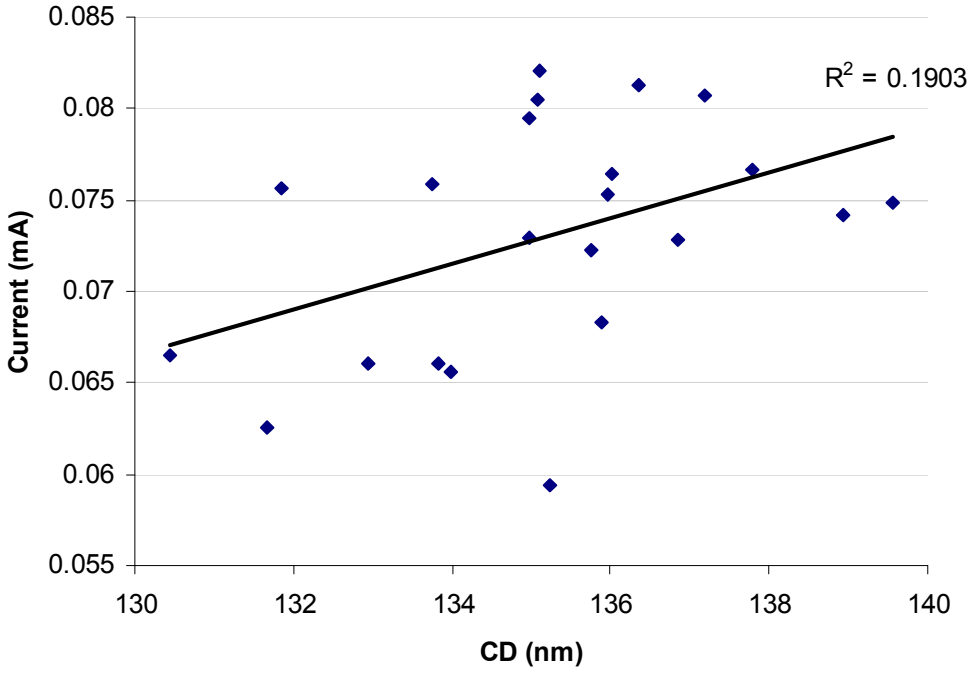

Figure 6.15. Plot of device saturation current through a standalone MOS device (at V_g = 1V) vs. average CD for each of the 22 test chips, for the vertical orientation.

The weakness of these relationships is very likely due to the wide variability of an individual MOS device; since only one standalone device is instantiated in each orientation for a given test chip, the measurement result is extremely noisy. To improve our ability to determine the proportion of total electrical performance variability owed to CD variation, we must be able to characterize several more devices on each test chips using the MOS array.

6.3.2 Attempts to Characterize the MOS Array

Despite the success of SPICE simulations performed using a netlist extracted from the Cadence layout submitted to the foundry, our attempts to characterize the physical MOS array parts were entirely unsuccessful. In running the scripts designed to cycle through the array of transistors and measure the properties of each, it became apparent the isolation of a single MOS cell through addressing could not be accomplished. Even worse, it was determined that applying even a small voltage—much less than the designed-for supply voltage of 1V—to the "supply" node while holding all other pads at zero voltage resulted in drastically excessive current being drawn (greater than the 1mA that is the maximum measurable current range for the tool due to compliance). That is, simply powering the circuit with the intended supply voltage (needed to enable the control logic) initiated a level of current that would drown out any possible measurement on a single transistor (max current ~70μA), regardless of the voltage levels placed at any other access nodes. A similar problem was seen when applying any voltage to the "gate" node while holding all other nodes at 0V—excessive, unexplained current.

During either of these applied-voltage "configurations," virtually no current should flow through the test structure, since several signals must propagate through the logic in a sequential nature to grant any access to the actual MOS devices and with these configurations, the control logic is disabled. At most, the only current measured should be the leakage current of a single row or a single column of transistors. Significantly, measuring the current at both the ground node and at the access node being used to apply the non-zero voltage ("gate" or "supply") yields equal and opposite levels of excessive current, indicating that alternate, unintended paths between these access pads and the ground node exist rather than a separate path (e.g., through the array itself) that is "opened" by the application of a voltage to the targeted access pad. In other words, it seems that current was being shunted away from the MOS devices and through an alternate path between the given access node and ground.

As shown in **Fig. 6.16**, the IV-characteristics of these alternate paths are somewhat confusing; for both the "gate" and "supply" nodes, the IV curves appear to be the output of cross-coupled diodes. However, the only two diodes in the circuit behave as we would expect (as evidenced by the reverse-biased diode between the "pulse" node and the ground node, which is accurately represented by the "pulse" IV curve in **Fig. 6.16**). Note that this reinforces the intuition that these IV curves could not be indicative of several devices from the MOS array acting in parallel; as the "gate" node is swung to a negative voltage, an increasing current flows when MOS devices would only be turned off "harder" by a negative gate voltage.

Unfortunately, due to the pad-limited nature of our design, no intermediate access points to the circuit were included for characterizing a potential bug in the circuit.

Additionally, the limited access to the manufacturing process limits speculation as to whether the problem with the circuit lies strictly within the design space, or whether a layer could have been misprocessed. Although all the other structures in the test chip worked perfectly, which might rule out a misprocessing, several metal layers were dedicated to the MOS array to ensure that shorting to the other structures (particularly the embedded serpentine ELM structure) would not occur. Still, misprocessing is an unlikely scenario. In summary, our best efforts to identify the source of the circuit bug were unsuccessful. The most obvious course of action if further investigation were designed would be to resubmit the design with several built-in hardware debug characterization nodes to evaluate signals at intermediate nodes within the circuit.

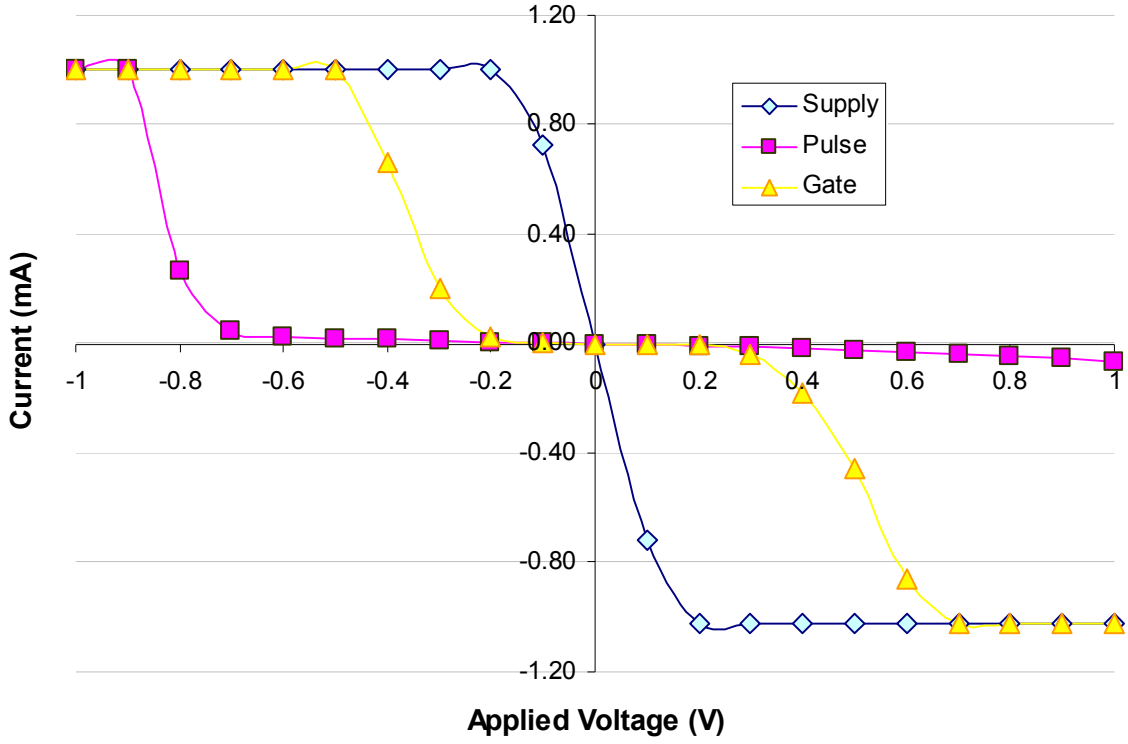

Figure 6.16. IV curves for three of the access pads when holding voltages at all other access points at 0V.

References

[6.1] K. Roy et al, "Estimation of Delay Variations Due to Random-Dopant Fluctuations in Nano-Scaled CMOS Circuits," *IEEE Journal of Solid-State Circuits*, Vol. 40, Issue 9, Sept. 2005, pp. 1787-1796.

[6.2] A. Bhavnagarwala, X. Tang, and J. Meindl, "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability," *IEEE Journal of Solid-State Circuits*, Vol. 36, No. 4, April 2001, pp. 658-685.

[6.3] R. Deaton and H. Massoud, "Manufacturability of Rapid-Thermal Oxidation of Silicon: Oxide Thickness, Oxide Thickness Variation, and System Dependency," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 5, No. 4, Nov. 1992, pp. 347-358.

[6.4] A. Asenov, S. Kaya, and J. Davies, "Intrinsic Threshold Voltage Fluctuations in Decanano MOSFETs Due to Local Oxide Thickness Variations," *IEEE Transactions on Electron Devices*, Vol. 49, No. 1, Jan. 2002, pp. 112-119.

[6.5] G. Schindler, M. Engelhardt, and M. Traving, "Size Effects and the Future of Interconnect Scaling," *VLSI Multilevel Interconnection Conference*, 2005, pp. 109-116.

[6.6] C. Yao et al, "Porous Ultra Low-*k* Process Technology Development for 65/45nm Nodes," *VLSI Multilevel Interconnection Conference*, 2005, pp. 37-39.

[6.7] Y. Shimizu et al, "Test Structure for Precise Statistical Characteristics Measurement of MOSFETs," *Proc. IEEE 2002 Int. Conference on Microelectronic Test Structures*, Vol 15, April 2002, pp. 49-54.

Chapter 7

Conclusions

7.1 Dissertation Summary

Based on the result of previously measured and newly designed, manufactured, and characterized test structures, this thesis has presented a complete analysis of spatial gate length variation. Such variation has been shown to consist of across-wafer, within-field, field-to-field, and pattern-dependent deterministic components, as well as a random, residual component that contains no significant additional spatial information. Several statistical models for spatial gate length variation of varying complexity were instantiated in an analytical, macromodel-based Monte Carlo simulation framework. This framework was used to demonstrate that the most complete and accurate statistical models (containing a full decomposition of sources of spatially deterministic variation) yielded substantially different predictions for circuit performance variability in comparison to simpler, "short-cut" statistical models. Moreover, the difference in predictions was shown to afford a designer more margin when using the more sophisticated spatial model: the projections under the most accurate model were more optimistic than those based on a simpler, "short-cut" model. The macromodel framework was also used to

show that the most efficient form of process control (between across-wafer and within-field process control) depends on the length of the typical critical path within the circuit. Thus, just as design choices should be process-aware, process control choices should be design-aware.

A second, industrial Monte Carlo simulation framework was used to evaluate the performance variability of canonical circuits implemented in various logic styles in the presence of device variation. It was shown that the susceptibility to variation was roughly equivalent across the various logic styles, with the passgate-based design style being slightly worse than either standard static or dynamic logic styles.

Finally, a second set of test structures was submitted for manufacture. Attempts to characterize the silicon parts were unsuccessful. Suggestions for a redesign are being generated.

7.2 Future Directions

Based on this work, several research directions can be pursued. In fact, two research projects have been launched during the final year of research for this dissertation. The first project aims to expand the complete spatial modeling of device variation from only the gate length parameter to the entire device. Additional test structures may be designed to specifically target the characterization of dopant distributions or oxide thickness variability. Within this project, an attempt will be made to identify simpler statistical models for spatial variation that do not sacrifice accuracy; although it has been shown here that a rigorous decomposition of variance accomplishes the most accurate statistical

representation of spatial process variation, the type of process characterization required for this analysis may prove to be too costly to implement.

The second project applies the spatial statistical models developed within this work as well as the models that will be developed within the first spin-off project described above in a framework that can be used to automatically optimize a circuit for critical path performance. By incorporating spatial models of variability, a group of critical paths can be optimized (by sizing device widths under an area constraint) to yield not only the best performance but also the least sensitivity to variation. Such optimization—referred to as statistical optimization—has been demonstrated to be superior to regular deterministic optimization, in which critical paths are optimized for performance alone with no attention paid to variability. Furthermore, it has been demonstrated that within the statistical optimization framework, using a model for spatial variation that contains complete decomposition of variance leads to a substantially improved optimization of circuit performance in comparison to using a "shortcut" model for spatial variation, such as the spatial correlation-only model [7.1]. This result confirms the ideas presented in Chapter 5 regarding the availability of extra design slack when using the more advanced and accurate spatial models.

In conclusion, it is our hope that this work will encourage a greater degree of attention to the spatially deterministic nature of device parameter variation. Although the costs of characterization may be imposing, the benefits of complete understanding of the spatial nature of variation can be substantial.

References

[7.1] Q. Y. Tang, P. Friedberg, G. Cheng, and C. J. Spanos, "Circuit Size Optimization with Multiple Sources of Variation and Position Dependant Correlation," *Design for Manufacturability through Design-Process Integration*, Proceedings of **SPIE** vol. 6521, pp. 108-109, 2007.

Appendix A

Autoprobe Programming

A.1 Basics of the Autoprobe UNIX Environment

For the bulk of use on the autoprobe, the Windows-based Metrics measurement toolkit is the software environment of choice. Metrics has a graphical user interface and straightforward command style, and is software that is commonly familiar to students through coursework. In this work, the Metrics measurement environment was used to characterize the ELM-only structures discussed in Chapter 3.

However, the Metrics environment has several limitations. First, Metrics is designed to facilitate low-level field-by-field testing; for a measurement routine that calls for a few measurements to be made for several die, Metrics is the optimal software. However, if a large volume of measurements is required for each die (as is the case for the MOS array designed and presented in Chapter 6), Metrics requires prohibitive time to compile the measurement code, which requires a several lines for each individual measurement. In addition, since a loop-based programming style is not build into Metrics, there is no easy way to perform clocked measurements or large volumes of sequential measurements.

Therefore, to characterize the MOS array, the UNIX Sunbase3 programming environment was used. The Sunbase3 environment consists of dozens of C files that can be compiled to provide a set of measurement routines that then may be called from the command line. Since for…to loops and the like are available in C, the Sunbase3 environment is convenient for processing a large number of measurements within a single die. However, due to the complexity of the C files themselves and their complete lack of documentation, there is a severe barrier to entry when one wishes to use the Sunbase3 environment. This Appendix will provide the Sunbase3 user with a brief introduction to the overall structure of the code, highlight the most important sections of code that the user may wish to edit, provide a list of base-level commands for setting/measuring voltages and currents, and share examples of code that was modified or written for the purposes of characterizing the MOS array.

A.2 Structure of the Sunbase3 Environment

The Sunbase3 environment consists of a set of files which are used to set up the equipment and describe the specifics of available test routines, and a pair of input files that are used to describe the wafer or test chip to be measured as well as to select the desired tests from the available routines. An overarching C file called *main.c* ties everything together, but need not ever be modified.

To set up the equipment, a large C file called *instruments.c* has been created. Nothing within this file needs to be adjusted from test to test with one exception: every time a probe card with a new configuration is used for testing, *instruments.c* must be updated to

contain the appropriate correspondence between the pins on the probe card and the values of the switch matrix. This is accomplished in a single line of code, at line 590:

```
   const int pin_convert[] =
{0,1,4,6,10,12,16,18,22,24,19,26,27,8,44,14,38,20,32,21,23,25,48,46,
42,40,36,34,30,28,2};
```

This line of code specifies the switch matrix values for each pin on the probe card, in this case from 1 to 15 along the top, and 16 to 30 along the bottom. The entire array must be preceded by a single zero to allow for a ground connection.

To set up the measurement routines, several files are used. The main workhorse in the routines will be a custom-named file such as *kelvmod.c* that contains commands to set voltages and currents, measure other voltages and currents, and save those measurements to a file. There are several such routine files that can either be modified or added to. In addition, the files *hash.c* and *modules.c* must be updated to contain header information and keywords for initializing and calling the various routines. Finally, each of the measurement routines, such as *kelvmod.c*, may refer to additional C files in the Sunbase3 directory such as *modtools.c*. The entire Sunbase3 framework is illustrated in **Fig. A.1**.

Finally, in order to call and execute the measurement routines, the files *prober.text* and *die.map* must be created. These two files are collected from within the Sunbase3 directory when the *main.c* file is executed, and provide the necessary information about measurement locations, chip layout on the wafer, and the desired measurement routines for each measurement site. The first of these files, *die.map*, gives the measurement locations within a single chip to be measured with respect to the initial placement of the probe tips. For example, the series of 9 DUTs from the micron-scale ELM test structure described in Chapter 3 is listed as:

```
8Ka 0,0
8Kb 0,-300
8Kc 0,-600
8Kd 0,-900
8Ke 0,-1200
8Kf 0,-1500
8Kg 0,-1800
8Kh 0,-2100
8Ki 0,-2400
```

The *die.map* also contains several fields that are used to initialize the measurement routines to be selected later; generally, though, any modification of these fields can be avoided by carefully setting up the measurement routine itself with the appropriate switch matrix values in mind.

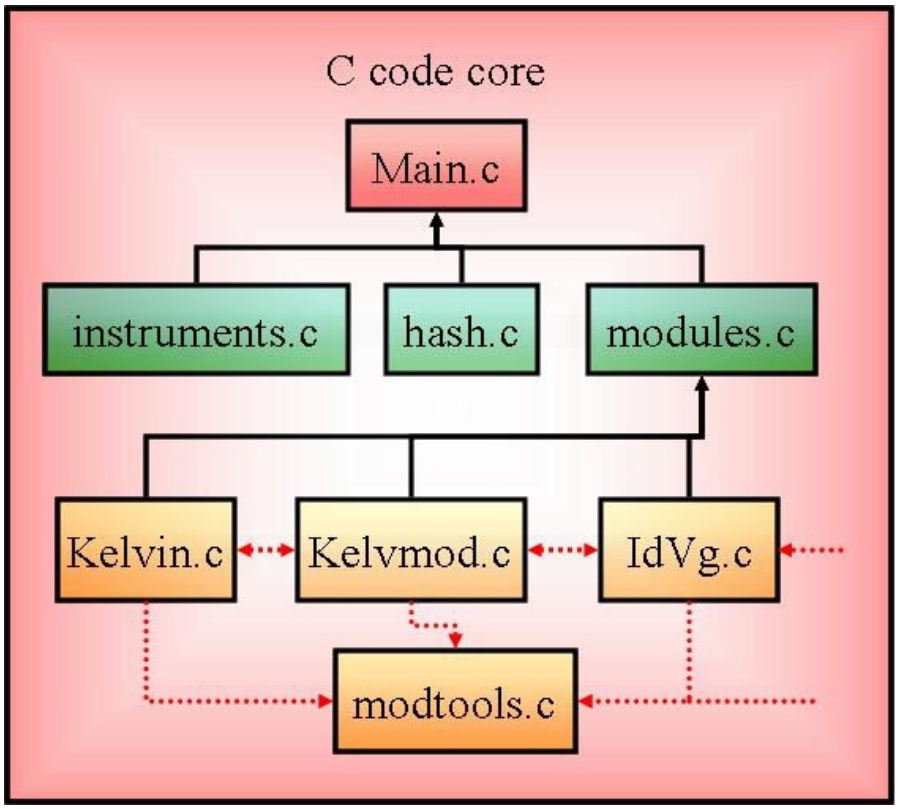

Figure A.1. Diagram of the Sunbase3 C code core environment.

The second file, *prober.text*, contains a grid of zeros and ones that denotes the chip layout on the wafer, with an *x* at the center to denote the first chip to be tested. For example, a sample wafer layout might look like:

```
0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0
0 0 0 1 0 1 0 0 0
0 0 0 0 0 0 1 0 1
0 0 0 0 0 x 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
```

For a single chip, the wafer layout is given by a single *x*. The *prober.text* file also contains a list of routines to be run at each selected chip.

A.3 Code for Characterizing the MOS Array

With the general Sunbase3 framework laid out, we now turn out attention to the construction of the actual measurement routines. There are three main commands used within this code: *connect*, *DCShold*, and *DCSmeasure*. The *connect* command connects a given SMU to a given pin of the probe card or ground. For example, *connect(2,0)* will connect SMU 2 to ground. The *DCShold* command will then connect a specified SMU to a voltage or current. For example, *DCShold(2,'V',1,0.1)* will set the voltage of SMU 2 to 1 volt (with compliance of 0.1V). Similarly, *DCShold(3,'I',0,1)* will hold the current of SMU at zero so that it can be used as a voltage probe. Finally, the *DCSmeasure* command performs a measurement of voltage or current at the specified SMU. For example, *dmake(DCSmeasure(3,'V'))-> value* will take a measurement of the voltage at SMU 3 and write it to the variable *value*.

To illustrate how these commands are strung together to compose a cogent measurement routine, the essential parts of the code used to characterize the MOS array described in Chapter 6 are presented here. Four clock functions, used to cycle the clocks on either the row or column vector of clocked elements, or to cycle a pulse into either the first stage or either that row or column, are listed below.

```
void clockEventCol(){
  connect (2,0);
  connect (2,16); /* connect to COL clock pad */
  DCShold (2,'V',1,0.1); /* open up FF to let pulse in*/
  usleep (10); /* pause to let signal get into FF */
  DCShold (2,'V',0,0.1); /* close FF to lock in pulse */
}

void clockEventRow(){
  connect (2,0);
  connect (2,17); /* connect to ROW clock pad */
  DCShold (2,'V',1,0.1); /* open up FF to let pulse in*/
  usleep (10); /* pause to let signal get into FF */
  DCShold (2,'V',0,0.1); /* close FF to lock in pulse */
}

void clockInPulseCol(){
  connect (1,0);
  connect (1,15); /* connect to pulse pad *
  DCShold (1,'V',1, 0.1);  /* set up the pulse */
  clockEventCol();
  DCShold (1,'V',0,0.1); /* end pulse */
}

void clockInPulseRow(){
  connect (1,0);
  connect (1,15); /* connect to pulse pad *
  DCShold (1,'V',1, 0.1);  /* set up the pulse */
  clockEventRow();
  DCShold (1,'V',0,0.1); /* end pulse */
}
```

These functions will be used to index a specific MOS transistor within the array. Next, once a single MOS device has been addressed, we wish to perform a short series of measurements, captured in the function *MeasureMOS*. The pads for the "source force," "drain force," and "gate" are connected to the first three SMU's, appropriate voltages are applied, and a measurement of the current through the device is performed. The

measurement is repeated for three different values of the gate voltage. Next, the "gate" and "source force" pads are shorted to V_{dd} and a 4-pt Kelvin type measurement is made using SMU's 3 and 4 to gather the source and drain voltages of the targets transistor. The four reported values are saved to variables which are passed to and from the function.

```
void MeasureMOS(float *i1_out, float *i2_out, float *i3_out,
float *Res_out){
      float v1, v2, dv, i_temp;
  connect (1,0);
  connect (2,0);
  connect (3,0);
  connect (1,20); /* connect to SourceForce pad */
  connect (2,21); /* connect to DrainForce pad */
  connect (3,22); /* connect to Gate pad */
  DCShold (1,'V',1,0.1); /* set SourceForce to Vdd */
  DCShold (2,'V',0,0.1); /* close DrainForce to Gnd */
  DCShold (3,'V',1,0.1); /* set Gate to Vdd */
  *i1_out = -1.0 * dmake(DCSMeasure (1,'I'))->value;
  DCShold (3,'V',0.7,0.1); /* set Gate to 0.7V */
  *i2_out = -1.0 * dmake(DCSMeasure (1,'I'))->value;
  DCShold (3,'V',0.4,0.1); /* set Gate to 0.4V */
  *i3_out = -1.0 * dmake(DCSMeasure (1,'I'))->value;

  connect(3,0);
  connect(4,0);
  connect (1,22); /* short Gate and SourceForce */
  DCShold (3,'I', 0, 1);
  DCShold (4,'I', 0, 1);
  connect(3,23); /* connect SMU3 to SourceSense */
  connect(4,24); /* connect SMU4 to DrainSense */
  v1 = dmake(DCSMeasure (3,'V'))->value;
  v2 = dmake(DCSMeasure (4,'V'))->value;
  dv = v2-v1;
  i_temp = -1.0 * dmake(DCSMeasure (1,'I'))->value;
  *Res_out = dv/i_temp;

}
```

Finally, the code which ties these functions together is contained in the routine *MOSArray_meas*. Within this routine, the following set of loops are executed:

1) a pulse is clocked into the column of flip-flops using *clockInPulseCol*
2) a pulse is clocked into the row of flip-flops using *clockInPulseRow*
3) the targeted MOS is measured using *MeasureMOS*
4) the measurement results are saved to file
5) the row of flip-flops is cycled to select the next MOS using *clockEventRow*
6) steps 3-5 are repeated for the total number of columns in the array

7) the column of flip-flops is cycled to select the next row of MOS using *clockEventCol*
8) steps 2-7 are repeated for the total number of rows in the array

In coded format, this measurement flow looks like:

```
void* MOSArray_meas (FILE *output) {
  int i, j, numRow, numCol;
  float i1_out, i2_out, i3_out, Res_out;
  numRow = 10;
  numCol = 450;
  i1_out = i2_out = i3_out = Res_out = 0.0;

  clockInPulseCol();
  for (i=2;i<=numRow;i++){
    clockInPulseRow();
    for(j=2;j<numCol;j++){
      MeasureMOS(&i1_out, &i2_out, &i3_out, &Res_out);
      fprintf (output, "%g\t%g\t%g\t%g\t(%d %d)\n", i1_out,
      i2_out, i3_out, Res_out, i, j);
      clockEventRow();
    }
    clockEventCol();
  }
}
```

To save the collected data for each test structure site, the following code is used:

```
  static char cmd[80];
  static char localfilepath[256];
  static char remotefilepath[256];
  static char timestr[80];
  static char logtimestr[80];
  char timefmt[80];
  char logtimefmt[80];
  time_t time_tNow;

  fflush(output);
  printf("Archiving...\n");
  strcpy(timefmt, "%y%m%d_%H%M%S");
  strcpy(logtimefmt, "%Y/%m/%d_%H:%M:%S");
  printf("getting current time...\n");
  time(&time_tNow);
  strftime(timestr, 80, timefmt, localtime(&time_tNow));
  strftime(logtimestr, 80, logtimefmt, localtime(&time_tNow));
  sprintf(localfilepath, "MOSArray_result_%s.xls", timestr);
  sprintf(cmd, "cp output %s", localfilepath);
  system(cmd);
}
```

In short, the data stored to the file *output* in the MOSArray_meas main block is appended with a timestamp, and is copied to a file named *MOSArray_result_*timestamp*.xls*. Finally, by comparing the timestamps with the order of measurement site executed, the data can be connected to the measurement site in the future.

References

[1.1] UC Berkeley Microlab's Autoprobe User Manual, http://microlab.berkeley.edu/labmanual/chap8/8.05.html.