A General Methodology for Assessing and Characterizing Variation in Semiconductor Manufacturing

by

Brian E. Stine

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

© Massachusetts Institute of Technology, 1997. All Rights Reserved.

Author ..........................................................

Department of Electrical Engineering and Computer Science
September 24, 1997

Certified by ..........................................................

Duane S. Boning
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by ..........................................................

James E. Chung
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ..........................................................

Arthur C. Smith
Chairman, Department Committee on Graduate Studies
Department of Electrical Engineering and Computer Science
Abstract
As device and interconnect dimensions continue to scale toward tenth-micron dimensions, maintaining process uniformity and compensating for deeply confounded interactions at each processing step is increasing in importance and complexity. While the impact of variation on circuits designed and fabricated using previous process technologies based on one micron or larger features has been relatively minor, as the feature size has approached quarter-micron dimensions and the wafer size has grown to 200mm and industry takes the next step to 300mm, the detrimental impact of variation has begun to assume a more prominent position. This is mainly attributed to a shrinking of feature dimensions without a corresponding decrease in machine precision and tolerance. “Worst-case” based approaches to coping with variation are often inadequate or highly conservative especially for key parameters such as the polysilicon critical dimension and the intra-level dielectric thickness which have a direct impact on circuit performance. The central focus of this thesis is to develop the framework for a general methodology (particularly for systematic pattern dependent or spatial variation) for assessing and coping with variation in semiconductor manufacturing as opposed to merely bounding the variation in a “worst-case” approach.

In this thesis, a novel methodology is developed for characterizing variation in semiconductor manufacturing with special emphasis on pattern dependent and spatial variation. New research contributions are specific methods for identifying the nature and scope of systematic pattern dependent and spatial variation, modeling the variation as a function of layout and/or process parameters, and determining the impact of pattern dependent variation on circuit performance. Also, a novel methodology for optimizing metal-fill for reducing ILD thickness variation in oxide CMP processes is developed as an example of variation reduction. Illustrations of this methodology are drawn from key interconnect and
device variation concerns with particular concern for ILD thickness variation in oxide CMP processes.

Thesis Supervisor: Duane S. Boning
Title: Associate Professor of Electrical Engineering and Computer Science

Thesis Supervisor: James E. Chung
Title: Associate Professor of Electrical Engineering and Computer Science

Co-Advisor: Edward R. Equi
Title: Manager, Process Integration, Digital Equipment Corporation
Acknowledgments

Pigmæi gigantum humeris impositi plusquam ipsi gigantes vident
(Pigmies placed on the shoulders of giants see more than the giants themselves).

Didacus Stella in Lucan, 10, tom. ii.

I do not know what I may appear to the world; but to myself I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.


Parting is such sweet sorrow
William Shakespeare
“Romeo and Juliet”, Act I, Sc. 2

This document represents the efforts and contributions of many, many people. First and foremost, I would like to acknowledge the support and guidance of my advisors Duane S. Boning and James E. Chung who made this project happen in the first place. Ted Equi, my co-advisor from Digital Equipment Corporation, has also been very instrumental in this project providing valuable feedback as well as contributing data and research effort. Jacob White, who served as a reader, was also the source of many good ideas especially for circuit impact methodologies.

This work has been funded by DARPA under contracts #DABT-63-95-C-0088 and #DAAH01-96-C-R904; AASERT grant #DAAHA04-95-I-0459 and an Intel Foundation Fellowship. I am extremely grateful for their financial support.

No research project in semiconductor manufacturing can be done without the successful collaboration of industry. The following list of people and organization have contributed time, resources, expertise, and/or guidance to this project

- Hans Eisenmann, Hendrik Schneider, Thomas Waas, and others from AISS, Gmbh
- Peter Burke, Kai Yang, and others from AMD
- Crid Yu, Costas Spanos, and others at the University of California, Berkeley
- David Avril Bell, Larry Binder, John Kitchin, and others from Digital Equipment Corporation
- Dirk Bartelink, O. Sam Nakagawa, Phil Nikkel, and Soo-Young Oh, Gary Ray, and others from Hewlett Packard
- Howard Landis, David LaPotin, and Sani Nassif from IBM
- Mark Bohr, Melton Bost, Sean Cunningham, Steve Duvall, Paco Leon, Peter Moon, Thye-Lai Tung, and Leo Yau from Intel Corporation
- Thomas Bibby, Randy Harwood, and Dennis Shy from IPEC/Planar
- Michael Berman, Ashok Kapoor (now with National Semiconductor), Bill
Loh, Moorthy Muthukrishnan, Dan Towery (now with Allied Signal), Sharad Prasad, and others from LSI Logic
• Karl Robinson and Scott Miekle from Micron Technology
• Anantha Chandrakasan, Thucydides “Duke” Xanthopoulos, and others at MIT
• Kathleen Perry (now with EDC, Corp.), Subramoney Iyer (formerly), and others from Motorola
• Peter Renteln from National Semiconductor
• Dennis Ciplickas, John Kibarian, and the rest of the gang at PDF Solutions, Inc.
• Larry Camilletti (formerly with Digital Equipment Corporation) at Rockwell Semiconductor Systems.
• Mike Fury formerly from Rodel
• John Curry from Strasbaugh
• Dale Hetherington, Tom Hill and Krishna Achuthan at Sandia National Laboratories,
• Jason Schneir, David Keller, James See and Chris Zenner from KLA/Tencor
• Iqbal Ali, Stephanie Watts Butler, Mi-Chang Chang, Joe Davis, Harold Hosack, PK Mozumder, Sudipto Roy, Greg Shinn, Charlotte Appel, Harold Hosack, and others at Texas Instruments
• Subhas Bothra, Milind Weling, and others at VLSI Technology, Inc.

The numerous contributions of the members of the staff of Digital Equipment Corporation’s FAB-6 fabrication facility in Hudson Massachusetts, Hewlett Packard’s ULSI Fabrication facility in Palo Alto, CA, LSI Logic’s Pilot Fabrication Facility in Santa Clara, CA, Sandia National Laboratories’ Microsystems Development Laboratory, and the Research and Development Fabrication Facility at Texas Instruments in Dallas, TX are acknowledged.

I would also like to thank Dawn Fitzpatrick at PDF Solutions, Inc. for putting up with my last minute requests and requisitions. The assistance of Scotti Fuller, Kevin Meyer-Golden, and Carol Nicolora is also acknowledged. I am also very grateful for the assistance and camaraderie of a multitude of fellow students at MIT past and present -- Mark Armstrong, Roxann Blanchard, Eric Chang, Steve Dao, Steve Decker, Rajesh Divecha, Jeff Gealow, Aaron Gower, Wenjie Jiang, Iliana Fujimori, Gary Hall, Keith Jackson, Abraham SeokWon Kim, Jee-Hoon Yap-Krska, Huy Le, Minh Le, Jen Lloyd, Joe Lutsky, Vikas Mehrotra, Charles Oji, Dennis Ouma, Tae Park, Melanie Sherony, Taber Smith, Dan Sobek, Tamba Gbondo-Tugbawa, and many more that I have unintentional missed.

Last, but certainly not least, I would like to thank my parents, brother, and family for supporting me and allowing me to start this project not to mention complete it.
# Table of Contents

1 Introduction ................................................................................................................ 19  
  1.1 The Modern Fabrication Facility ..................................................................... 19  
  1.2 Statistical Metrology ....................................................................................... 20  
  1.3 Sources of Variation ........................................................................................ 22  
  1.4 Statement of the Problem ................................................................................. 25  

2 Variation Assessment ................................................................................................. 29  
  2.1 Introduction ...................................................................................................... 29  
  2.2 Variation Classification And Definition .......................................................... 31  
  2.3 Wafer-level Variation Extraction ..................................................................... 33  
    2.3.1 Downsampled Moving Average Estimator (DSMA) ............................ 34  
    2.3.2 Meshed Spline Method (MSM) ............................................................. 37  
    2.3.3 Regression Based Estimators ................................................................. 39  
    2.3.4 Combined Techniques ........................................................................... 40  
    2.3.5 Simulation Results ................................................................................. 42  
    2.3.6 Parameter Selection Guidelines for the DSMA and MSM Estimators .. 43  
    2.3.7 Discussion of Wafer-level Variation Extraction Techniques ................ 46  
  2.4 Die-level Variation ........................................................................................... 47  
    2.4.1 FFT Based Method ................................................................................ 47  
    2.4.2 Simulated Dataset .................................................................................. 50  
  2.5 Wafer-die Interaction Terms ............................................................................ 50  
    2.5.1 Spline Based Method ............................................................................. 51  
    2.5.2 FFT Based Method ................................................................................ 53  
    2.5.3 Simulated Extraction and Error Analysis .............................................. 57  
  2.6 Residual Analysis ............................................................................................. 59  
    2.6.1 Correlogram Methods ............................................................................ 59  
    2.6.2 Die Mean Residual Test ......................................................................... 60  
    2.6.3 Simulation Results ................................................................................. 62  
  2.7 Experimental Dataset ...................................................................................... 63  
  2.8 Summary .......................................................................................................... 67  

3 Rapid Variation Assessment and Diagnosis .............................................................. 71  
  3.1 Introduction ...................................................................................................... 71  
  3.2 Wafer-Level Variation ..................................................................................... 71  
  3.3 Die-Level Variation ......................................................................................... 73  
  3.4 Wafer-Die Interaction Terms ........................................................................... 74  
  3.5 An Example ..................................................................................................... 74  

4 Variation Modeling: Investigative Modeling ............................................................. 81  
  4.1 Introduction ...................................................................................................... 81  
  4.2 Motivation ........................................................................................................ 83  
    4.2.1 Experimental Methodology ................................................................... 83  
    4.2.2 Wafer level model .................................................................................. 84  
    4.2.3 Die level and Wafer-Die Interaction Term Model ............................... 85  
    4.2.4 Poly-CD Data Analysis ....................................................................... 85  
    4.2.5 Summary ............................................................................................... 90
List of Figures

Figure 1.1: The three metrics of product quality. Ideally a product should score high on all three axes, but cost considerations often make this unobtainable. .................................20

Figure 1.2: Variation can often be classified into logical atoms such as lot-to-lot and wafer-to-wafer variation. Variation can be spatially driven or temporally driven, but time and space are often stand-ins for more physically rooted phenomenon. ........................23

Figure 2.1: Variation decomposition method flow diagram ..............................................32

Figure 2.2: Downsampling moving average method (a, b) input sequence, (c, d) moving average window, (e, f) downsampler, (g, h) result of averaging and downsampling the input sequence, (i, j) final result after interpolation. (b, d, f, h, j) is the frequency response of (a, c, e, g, i) ..............................................................................................................................36

Figure 2.3: Meshed spline method (a, b) the input sequence, (c, d), the result of striping in the y direction, (e, f), the result of striping in the x direction, (g, h) the MSM estimate for the wafer-level variation shown in (a). (b, d, f, h) is the frequency response of (a, c, e, g) . 38

Figure 2.4: Artificial data set (described in detail in Appendix A) ....................................42

Figure 2.5: Autocorrelation function for artificial data set ................................................45

Figure 2.6: (a,b) A sample signal and Fourier transform used to evaluate the accuracy of (2.8). (c) Estimates of X[0] drawn from 100 independent trials. .................................49

Figure 2.7: Effect of degrees of freedom on smoothness ..................................................52

Figure 2.8: Exponential spar method .................................................................................53

Figure 2.9: (a) The wafer-level, die-level, interaction, residual, and raw variation components for an illustrative one-dimensional wafer. (b) the frequency spectra of the known interaction terms versus the die-level (or averaged die-level) residuals for (b) a small residual component, (c) a large residual component, and (d) a large residual component but averaged across many wafers. ..................................................................................................56

Figure 2.10: Wafer-die interaction terms for an artificial data set. Extracted using spline based estimator ...........................................................................................................58

Figure 2.11: (a) Quantile-quantile plot, (b) correlogram function for total residuals, and (c) die mean residual test from artificial data set .......................................................61

Figure 2.12: HP Mask Design Showing (a) die layout (area intensive structure lower right),
(b) subdie layout: four capacitors with layout factors near local linewidth measurement structures, and (c) capacitive ILD thickness test structures with (d) layout experimental design factors

Figure 2.13: Wafer-level variation, die level variation, interaction terms, and total residuals for a die near the center from the HP Mask Set

Figure 2.14: Wafer-die interaction variation comparison for HP mask set (a) for one die near center (b) for one die near edge (wafer B5)

Figure 2.15: Die Mean Residual Test for Wafer M2

Figure 2.16: Histogram Comparison and Quantile-Quantile Plot for HP Mask Set (Wafer B5)

Figure 3.1: The single structure method samples the same structure on each die across all dies on a wafer typically requiring 75-100 samples per wafer. Most automated metrology tools can measure a wafer using this strategy in about five minutes

Figure 3.2: The wafer-level variation estimator using the single structure method (SSM) can yield erroneous results especially if a site is selected which does not modulate with the wafer-level variation

Figure 3.3: A rapid estimator for the die-level variation can be computed using position averaging. Position averaging is computed by computing the average of all measurements for the same structure across all die and then subtracting the grand mean

Figure 3.4: Raw ILD Thickness variation

Figure 3.5: The wafer-level variation estimated using the DSMA technique (a,c) and the SSM technique (b,d). The extracted wafer-level variation is shown for one wafer from each pad split

Figure 3.6: The die-level variation estimated using the FFT techniques (b,d) and using the position averaging techniques (a,c). Data is shown for a representative wafer from each pad split

Figure 3.7: (a) A comparison of the single-structure method versus the down-sample moving average method and (b) the position averaging method versus the FFT method

Figure 3.8: The wafer-die interaction terms (a, b) extracted using the FFT techniques discussed in Chapter 2. The estimated 1 + a surface for a wafer from each pad split is shown in (c, d)

Figure 4.1: Variation modeling relies upon the results of the variation assessment phase. There are two types of variation modeling approaches: investigative modeling and analy-
ical modeling. Often the results of an investigative model will form the basis of or lead to the development of an analytical model .................................................................82

Figure 4.2: The VarNOVA procedure .................................................................................85

Figure 4.3: Raw Polysilicon Critical Dimension Variation for Wafer 17 .........................86

Figure 4.4: Several Extracted Wafer-Level Variation Components ...................................87

Figure 4.5: The Spatial Dependence of the Model Coefficients for Wafer 21, Lot “A” ...88

Figure 5.1: The CMP Characterization Masks: the area, pitch, density, and aspect ratio masks are shown in Figures 5.1a-d respectively. Each die is 12mm x 12mm .........................94

Figure 5.2: The short-flow process flow used in the pad comparison study. .......................98

Figure 5.3: ILD thickness versus structure area, pitch, pattern density, and aspect ratio for each pad type. Note that for the area and aspect ratio data, Δ ILD thickness values are listed since the effect of pattern density has been removed from the data .........................100

Figure 5.4: The 3-D profilometry measurements for each mask across each pad type. The profilometry measurements have been leveled using the combined optical/profilometry techniques ............................................................101

Figure 5.5: A simplified example to aid in the definition of pattern-density. .................103

Figure 5.6: The definition of planarization length. Typical planarization lengths are on a scale of several mm .................................................................104

Figure 5.7: The definition of interaction distance shown on an example layout ..............105

Figure 5.8: Estimated slope of ILD thickness versus pattern density computed at different interaction distances from 3 to 4.5mm for the density mask (a) and for the area mask (c). For comparison the maximum $R^2$ method is shown in (b) and (d) for the density mask and for the area mask respectively .................................................................108

Figure 5.9: Profilometry measurements for a center die from the area mask polished with an IC-1400 pad (b) versus the predicted ILD thickness (a) based on computed pattern-density. Discrepancies between the model and profilometry measurements are due to lack of knowledge about the window shape, uncorrected stage bias and wafer warping, and second order effects .................................................................111

Figure 5.10: A comparison of the dependence of ILD thickness on perimeter/area (a) versus aspect ratio (b) for sample wafers from each pad type. The effect of pattern density has been removed to show only the perimeter/area or aspect ratio dependence .................................115
Figure 5.11: ILD thickness versus separation distance for the aspect ratio mask. Each data point represents a structure with the same aspect ratio but different separation distance. The density dependence has been removed to isolate structure separation effects..............116

Figure 6.1: Originally statistical simulation techniques were used to understand the impact of process variation on circuit performance. This approach is successful and accurate for parameter variation which has a strong random component. For strong systematic components, the methodology presented here based on deterministic prediction of variation is more appropriate.................................................................120

Figure 6.2: Estimation of the effect of systematic sources of parameter variation on circuit performance is hampered by a tendency of certain forms of systematic variation to be localized to several microns while other forms tends to assume more global dependencies which are outside the bounds of the electromagnetic region of interest in a simulation. In this way, TCAD types simulations alone are not sufficient.................................121

Figure 6.3: The Hopkins model used in aerial imaging simulators. Figure and notation derived from [77]..................................................................................................................124

Figure 6.4: The “ideal” as-drawn features from an SRAM cell and aerial imaged features . 125

Figure 6.5: The methodology for estimating the impact of systematic pattern dependent variation on circuit performance. In this methodology, key elemental steps include meshers, layout-parasitic extraction engines, 3-D solvers or analytical capacitance equations, and variation models........................................................................................................129

Figure 6.6: The connectivity of a layout is usually represented as a graph for each electrical net (in this case P1 and P2 nets). Each node in a graph represents a polygon in the layout and the weights between nodes represents the way nodes connect to neighboring nodes. ... 131

Figure 6.7: The nets of interest as well as the net “halo” are segregated from the layout. In this way, the critical nets of interest can be simulated in 3-D solvers.........................132

Figure 6.8: An example of rectilinear gridding (left) and Delaunay gridding (right). Two example simulation windows and overlapping regions are also shown..................133

Figure 6.9: This figure shows a sample p ladder distribution scheme for distributing resistors and capacitors within a simulation window. In this figure, the node count numbers (n1, n2) are 6 and 3. In this figure, the “halo” geometries are not shown............................135

Figure 6.10: A simplified balanced H-bar clock tree. Since the length of the line is the same from the buffer (at the center of the layout) to the end of each path and a ground plane runs under the clock line, the skew should ideally be zero. However, the pattern density underlying the clock line changes, and the realistic clock perform will skew to some degree con-
strained by the amount of ILD thickness variation .............................................................. 137

Figure 6.11: The estimated ILD thickness variation for the balanced H-bar clock tree in Figure 6.10. The ILD thickness variation model in [59] was used to estimate the amount of ILD thickness variation. If the ILD thickness was completely uniform, a balanced H-bar clock tree should be skew free; however, ILD thickness variation perturbs the balance of the tree resulting in some clock skew .................................................................................. 138

Figure 6.12: The results of simulating the balanced clock tree. The ILD thickness variation and linewidth is shown in (a) and (b). The resistance and capacitance values are shown in (c) and (d) respectively. The line was divided into 100 micron lumps so capacitance and resistance is reported in units per 100 microns of length. The final simulated output voltages for two of the paths is shown in (e). ............................................................................ 139

Figure 6.13: A 256Kx1 SRAM array implemented in a 0.50 micron three level metal process. The array is constructed from sixteen 16K macro-cells plus decoder circuitry along the central region ............................................................................................................. 140

Figure 6.14: The ILD-1 (between metal-1 and metal-2) and ILD-2 (between metal-2 and metal-3) thickness variation for the 256Kx1 SRAM array shown in Figure 6.13. The total range is approximately 0.2 μm in each case .................................................................... 141

Figure 6.15: The estimated metal-CD variation for one of the metal layers in Figure 6.13. Since the stepper leveling contribution is random from field to field, this figure represents only one of a family of delta-CD contours that are possible ........................................... 142

Figure 6.16: A 100 μm by 10 μm slice from the 256Kx1 SRAM array showing the bit-line surrounded by the halo structures above, below, and to either side. Structures below Metal-1 (such as poly and active area) did not couple significantly to the bit-line structures; thus, they were removed .......................................................................................................... 143

Figure 6.17: A DOE and macro-model was generated of the bit-line capacitance to ground and halo (i.e. with the halo grounded). This model allows for a rapid conversion between variation in these five physical parameters to bit-line capacitance ................................................................. 145

Figure 6.18: The capacitance of a bit-line near the edge and near the center to ground and the halo structure (held grounded) for the 256Kx1 SRAM array. Significant skew can be seen for bit-lines near the edge versus bit-lines near the center. Also, bit-lines near the edge show more variance compared to near the center .................................................................................. 146

Figure 6.19: A 64x8 SRAM macro-cell. The layout is approximately 180 μm x 140 μm ................................................................................................................................. 147

Figure 6.20: The estimated poly-CD variation across all coordinates in the cell. Aerial imaging simulation was used to estimate the poly-CD variation .............................................................................. 148

Figure 6.21: (a) Drain current versus drain voltage for Vgs at the supply rail. (b) Saturation
or drive current versus channel length for the n-mos and p-mos devices......................149

Figure 6.22: Simulation of the write-access behavior for the 64x8 SRAM macro-cell with poly-CD variation enabled. Significant skew is seen across all the data lines for this particular address ..................................................................................................................................150

Figure 6.23: Delay versus input address for the 64x8 SRAM cell. The maximum skew across all output lines depends on the address chosen.............................................................151

Figure 6.24: For simultaneous switching conditions, the amount of skew increases dramatically for the 64x8 SRAM cell. The exact cause for this behavior is unknown. .............152

Figure 7.1: Range of ILD thickness variation versus die size. As the die size has increased, the range of ILD thickness variation has also increased. This can be attributed to the size of the chip approaching and then surpassing the finite planarization length of the CMP process. The data shown was simulated using the model in [89]. The same layout was used for each data point and scaled appropriately. Although this would seem to indicate that scaling should help the problem, in reality, scaling forces are checked by ever increasing demands for higher transistor counts leading to larger and larger die sizes. ........................................156

Figure 7.2: Dielectric thickness distributions between metal-1 and metal-2 (a, b) and between metal-2 and metal-3 (c, d). Reticle “A” did not have metal-fill added while reticle “B” had metal-fill patterning. ..........................................................................................................................161

Figure 7.3: A contour plot of the achievable pattern density inside of an empty 1mm x 1mm area versus the metal-fill design rule. The metal-fill patterning scheme is shown in the inset. The buffer distance for this example has been fixed at 25μm. A minimum linewidth of 0.35μm and a minimum linespace of 0.45μm has been assumed.................................163

Figure 7.4: An illustration of vertical line filling (a) which is recommended for grounded metal-fill and block filling (b) which is recommended for floating metal-fill. Note the bridging bars that have been added in (a) to permit a connection to ground. Because the fill in (a) is grounded the metal-fill lines do not permit coupling between regions A and B although there may be an overall increase in capacitance and delay.................................166

Figure 7.5: A stylized illustration depicting the definitions and capacitance typically of interest in (a) floating metal-fill with block patterning and (b) grounded metal-fill with vertical line patterning.............................................................................................................169

Figure 7.6: Contour plots of 100(C* - 1) are superimposed on contour plots of constant pattern-density to form a design chart for the coupling and optimizing of pattern-density/uniformity specifications with interconnect capacitance specifications......................................................169

Figure 7.7: The metal 1 layer of the layout described in [89] used in the metal-fill case study discussed in Section 7.5. The layout is about 7.9mm by 9.2mm and the minimum linewidth and space at metal 1 is 3μm.................................................................172
Figure 7.8: a) Interconnect capacitance per unit length for two parallel lines 0.6 micron thick. Note the relative independence of capacitance to linewidth. (b) The effect of interconnect capacitance for two parallel lines versus spacing and metal thickness with the metal width fixed at 2.5 μm. Note the relatively good agreement ($R^2 > 0.99$) between the simulated capacitance values and the model shown and discussed in Section 7.5. (c) The simulation structure used in the comparison shown in (a,b). .................................................. 174

Figure 7.9: The simulation structure used to evaluate the capacitance approximation in (7.4). For the comparison, all lines except b1 were grounded, and the capacitance between b1 and every other metal region was simulated using a three dimensional capacitance solver. ..................................................................................................................................... 175

Figure 7.10: Design charts of minimum pattern density constraints superimposed on a relevant metric for evaluating the effect of a particular metal-fill design rule on interconnect capacitance. Design charts are shown for buffer lengths of 10μm, 25μm, 50μm, and 100μm. These charts allow one to contrast uniformity constraints with capacitance requirements. In reality, many of the negative capacitance numbers shown are slightly larger due to neglected fringing elements. ..................................................................................................................................... 177

Figure 7.11: Detailed view of the metal-fill design chart for a buffer length of 25μm. For the case study discussed in Section 7.5, the optimal choice of 9μm and 3μm for linewidth and linespace is highlighted. This value achieves a minimum pattern density of 50% and minimizes any added capacitance associated with the metal-fill. Linewidths and spaces below 3μm were not considered manufacturable. ..................................................................................................................................... 178

Figure 7.12: The simulated ILD thickness variation based on the model presented in [89] for the metal-1 layer of the layout shown in Figure 7.7 without metal-fill (a,b) and with metal-fill (c,d) with a buffer-length of 25μm, a block width of 9μm, and a block spacing of 3μm ................................................................................................................................. 180

Figure C.1: An illustrative diagram used in the calculation of the fringing capacitance formula discussed in Section 7.5 ..................................................................................................................................... 194
List of Tables

Table 2.1: Comparison of the Efficiency of the Wafer-Level Estimators ................................................. 43
Table 2.2: Effect of Downsampling on DSMA Estimator Accuracy ......................................................... 46
Table 2.3: Model Dataset: Die Level Variation Extraction Results ......................................................... 50
Table 2.4: Simulated Wafer-Die Interaction Term ...................................................................................... 58
Table 2.5: Extracted Total Residuals ........................................................................................................ 62
Table 4.1: Transistor Layout Factors ....................................................................................................... 84
Table 4.2: ANOVA Table for Wafer 21, Lot “A”, Die 44. ...................................................................... 88
Table 4.3: The Correlation Coefficient for the Generated VarNOVA Model ........................................ 89
Table 5.1: The Polishing Parameters Used in the Pad Comparison Study .............................................. 97
Table 5.2: ILD Thickness Variation Models for the Pad Experiment ...................................................... 109
Table 6.1: Several different components of systematic and random lithography variance and the estimated contribution of each factor to the total lithography variance [76]. .......................................................... 123
Table 6.2: The DOE experiment used to construct a macro-model of bit-line capacitance versus metal linewidth and ILD thickness values ........................................................................................................ 144
Table 7.1: The ANOVA table for the ILD-1 level for the first experiment discussed in Section 7.2. The standard deviation is modeled as a function of the wafer, the die, and the reticle used. Only the reticle effect (metal-fill patterning versus no metal-fill patterning) was found to be statistically significant. ........................................................................................................ 160
Table 7.2: Comparison of the capacitance approximation in (7.4) versus simulation. ......................... 176
Table 7.3: Buffer Length, Block Width, Block Space values which minimize the effect of the added metal-fill pattern on interconnect capacitance and achieve a minimum pattern density value of 50% (extracted from Figure 7.10). .......................................................... 179
Table A.1: Die-Level Variation Data (One Die Shown) ......................................................................... 188
Chapter 1

Introduction

1.1 The Modern Fabrication Facility

This work is concerned with semiconductor manufacturing processes such as those need to make microprocessors, memory, and other integrated circuits. Semiconductor manufacturing is conducted in fabrication facilities called “fabs” for short. Each fab typically costs $500 million to $2 billion dollars in capital expense plus an enormous overhead expenditure in people and materials. Fabs are the most expensive manufacturing plants in existence, and although the technology used to fabricated integrated circuits is almost identical from fab to fab, there is significant variation in the yield, even for the same design, from fab to fab [1]. Individual products, or chips, are manufactured on wafers containing anywhere from 30 to 200 chips. Typically a wafer will see approximately 400 processing steps before completion and perhaps 100 key parameters can be identified and measured. A failure in any one of these steps or excursions of any key parameter outside of specification often renders many chips or even entire wafers useless.

As shown in Figure 1.1, three metrics can be used to quantify the success of quality of a chip under manufacture: manufacturability, reliability, and performance. Manufacturability is associated with the throughput, control, uniformity, capacity, and yield of the product in the manufacturing line. Reliability is a measure of the robustness of the process or design, i.e. what is the expected lifetime of the product under normal operating conditions. Finally, performance is a measure of the speed of the product. The ideal product should score high on all three axes. Economic considerations are implicitly realized by these three metrics: scoring high on all three axes while technologically feasible may not be economically feasible.
Many factors play a role in determining where a product will map onto the manufacturability, reliability, and performance axis. Arguably, a chief factor is process variation, i.e. the tendency of a particular key parameter or process step during manufacture to vary, since excessive variation can either lead to outright failure or to soft failure modes (which map back into performance or reliability concerns). This work is focused on variation, particularly spatial (i.e. variation of key parameters from location to location on a wafer) and pattern-dependent (i.e. variation of key parameters from product to product). In this work, a general methodology is presented for assessing spatial and pattern-dependent variation (i.e. qualifying and quantifying the sources and nature of the variation), modeling variation, and determining the impact of the variation on circuit performance and/or product manufacturability. Statistical metrology is the body of methods used to accomplish these goals.

![Figure 1.1: The three metrics of product quality. Ideally a product should score high on all three axes, but cost considerations often make this unobtainable.](image)

1.2 Statistical Metrology
The term “statistical metrology” does not adequately describe what is meant. Obviously, statistical metrology is connected with statistics and with measurements and variation, but
it is more than that. While there is still significant debate over exactly what is meant or
encompassed by statistical metrology [2, 3], a popular definition is: “Statistical metrology
is the body of methods for understanding variation in microfabricated structures, devices,
and circuits” [3].

Statistical metrology can be disaggregated into three steps or areas: variation assess-
ment, variation modeling, and variation impact. In variation assessment, the primary goal
is to understand the source and extent of variation in the sense of answering the questions
where? and how much? In particular, where is the most amount of variation centered? Is it
intra-die, wafer-level, or wafer-to-wafer variation? How much of the total variation is
attributed to wafer-level, intra-die, wafer-to-wafer, etc.? In the variation modeling phase,
the primary goal is to model any systematic variation, either at the wafer-level or die-level,
as a function of either layout/pattern factors (in the case of die-level variation) or process-
ing parameters (usually in the case of wafer-level variation) through the use of traditional
or modified ANOVA (Analysis of Variance) or other statistical modeling techniques.
Finally, in the impact phase of statistical metrology, the primary concern is understanding
the relationships between parameter variation and the resulting circuit performance or
manufacturability and identifying methods for decoupling the impact or for variation
reduction.

In any statistical metrology approach, a number of tools and methods are involved.
Chief among these are usually a heavy use of statistics particularly design of experiments
(DOE), analysis of variance (ANOVA) and spatial statistics. TCAD simulations such as
gridders, meshers, and solvers are often involved. Signal processing and numerical algo-
rithms and design automation tools (EDA) such as layout engines and layout-parasitic
extraction engines are often involved. Finally, the use of in-line inspection/metrology tools
and the processing tools and in-situ monitoring or diagnosis tools cannot be neglected.
Statistical metrology methods form the basis for a wide variety of applications such as equipment characterization/selection; equipment, consumable, or process optimization or synthesis; process or equipment control; and layout optimization. In equipment characterization or selection, several different equipment types are evaluated (usually in some short-flow process loop) and the results of the evaluation are analyzed as part of a decision on which tool to purchase. Equipment, consumable, or process optimization involves evaluation of a large design space of the tool and consumables to determine which settings are optimum (usually in the sense of uniformity or variability). Process or equipment control uses variation models and information to make intelligent choices about machine settings to narrow or reduce the variation profile currently presented by the tool or process without control. Layout optimization uses variation models to determine a layout strategy for pattern-dependent variation reduction.

1.3 Sources of Variation
Variation in some physical or electrical parameter may manifest itself in several ways. A key distinction is systematic versus random constituents in the parameter distribution. An important goal is to isolate those systematic, repeatable, or deterministic contributions to the variation from a set of deeply confounded measurements. Without detailed understanding of the individual contributions, the distribution is typically considered to be “random” and large. An understanding of the contributions enables one to focus variation reduction efforts more appropriately, or to design the device or circuit to compensate for the expected variation.

Variation manifests itself across time and across space. In each case, the variation appears at a number of different scales: lot-to-lot, wafer-to-wafer, within-wafer, and intra-die (Figure 1.2). The separation of variation by unique signatures at different scales is a key element used to pry apart and analyze such variation. Also, an important requirement
is the design of appropriate measurement strategies which can take the scope of such variation into account.

**Lot-to-Lot**

**Wafer-to-Wafer**
(or within Lot)

**Within Wafer**

**Intradie**

Figure 1.2: Variation can often be classified into logical atoms such as lot-to-lot and wafer-to-wafer variation. Variation can be spatially driven or temporally driven, but time and space are often stand-ins for more physically rooted phenomenon.

Lot-to-lot variation is the tendency of the lot mean of a device or process parameter (e.g. the mean of channel length computed over the entire lot) to vary from one lot to the next. Lot-to-lot variation is often monitored using statistical process control, and may be compensated for using run by run or other feedback control approaches, e.g. [4]. Wafer-to-wafer variation may be either temporal or spatial in nature. Temporal wafer-to-wafer variation is generally caused by drift in process equipment operation from one wafer to the next. This variation is increasing in importance as single-wafer processing equipment expands in use. Spatial wafer-to-wafer variation may also result from non-ideal process equipment, e.g. due to different positions of wafers in a boat during a batch furnace step.
Wafer-level (or within-wafer) variation is generally caused by additional equipment non-uniformity and other physical effects such as thermal gradients and loading phenomena. Typically, wafer-level variation is low frequency and smooth, and neighboring points are likely to be highly correlated with each other. Also, wafer-level variation often exhibits symmetrical properties such as radial (or “bull’s eye”) patterns or slanted planes.

Die-level (or intra-die) variation is often caused by layout and topography interaction with the process. Key examples include pattern planarization in chemical mechanical polishing [5], and critical line-width dimension variation in channel length or metal lines [6]. Intra-die variation has only recently received appreciable attention, in part due to the need for a large amount of statistically meaningful data, and the prevailing belief that intra-die variation is inconsequential compared to lot-to-lot, wafer-to-wafer, and within-wafer variation. Several studies [5, 6, 7] have shown that this is not the case and that intra-die variation is often much larger or comparable to the other variational sources. Finally, variation at an even smaller scale, i.e. variation within individual features and devices is of interest especially for reliability concerns.

In this work we also identify a hybrid component termed the wafer-die interaction terms which capture the tendency of within-die variation to be accentuated or attenuated depending on location within the wafer. It may also represent a term proportional to the die-level variation modulated by the wafer-level variation. The majority of the methodology presented is focused on spatial and pattern dependent variation; thus, the wafer-level, die-level, and interaction terms are of special interest.

Time and space are generally “stand-ins” for more fundamental sources of variation. This is an important underlying principal in understanding variation. In lot-to-lot or wafer-to-wafer variation, for example, fundamental causes for equipment drift may be film buildup on chamber walls in plasma processes; both buildup and the resulting variation
signatures are related through a similar dependence on time. Similarly, a key component of intra-die variation is pattern dependency; understanding positional dependencies and correlations between intra-die variation patterns and layout patterns gives a clue as to the causal forces at work.

1.4 Statement of the Problem
As device and interconnect dimensions continue to scale toward tenth-micron dimensions, maintaining process uniformity and compensating for deeply confounded interactions at each processing step is increasing in importance and complexity. While the impact of variation on circuits designed and fabricated using previous process technologies based on one micron or larger features has been relatively minor, as the feature size has approached quarter-micron dimensions and the wafer size has grown to 200mm and industry takes the next step to 300mm, the detrimental impact of variation has begun to assume a more prominent position. This is mainly attributed to a shrinking of feature dimensions without a corresponding decrease in machine precision and tolerance. “Worst-case” based approaches to coping with variation are often inadequate or highly conservative especially for key parameters such as the polysilicon critical dimension and the intra-level dielectric thickness which have a direct impact on circuit performance. The central focus of this thesis is to develop the framework for a general methodology for assessing and coping with variation in semiconductor manufacturing as opposed to merely bounding the variation in a “worst-case” approach. This work is primarily concerned with systematic variation particularly spatial and pattern-dependent variation.

This thesis is divided into eight chapters. Chapters 2 and 3 discusses statistical analysis techniques for the variation assessment phase. In Chapter 2, methods for factoring variation into four main components as defined in section 1.3: wafer-level variation, die-level variation, wafer-die interactions, and residuals which are assumed to be orthogonal to the
systematic components are discussed. Several contributions of this research can be identified.

- Three different techniques for extracting the wafer-level variation: the down-sampled moving average, the meshed spline method, and a regression based approach.

- A Fast-Fourier transform (FFT) technique for extracting the die-level variation. This technique was originally proposed in [7], but this work builds on the original technique by presenting guidelines for the number of dies to measure on a wafer to get a given estimator accuracy and a discussion on the merits of the interpolation step, originally presented in [7], with respect to estimator accuracy.

- Two techniques for estimating the wafer-die interaction terms and parameter extraction guidelines for obtaining best results from these estimators.

- Several analysis tools and tests to determine the presence or absence of any systematic spatial variation in the residual component.

Chapter 3 also details contributions towards the decomposition of variation into wafer-level, die-level, and interaction term components, but which are meant for rapid analysis in situations where the rigorous sampling constraints necessary for the techniques presented in Chapter 2 are not practical or available. The contributions of this chapter are:

- The single structure method for rapidly extracting the wafer-level variation and a discussion of the limitations of this technique.

- A position averaging technique for estimating the die-level components. This technique was first discussed in [8], but we extend this technique by comparing the accuracy of the position averaging technique to the FFT based die-level technique presented in Chapter 2.

- A regression approach based on the assumption of a multiplicative model for estimating the wafer-die interaction terms.

Techniques for variation modeling are presented in Chapters 4 and 5. Chapter 4 presents techniques for investigative variation modeling (i.e. for situations where no physical or semi-empirical models are known to be valid) while the material offered in Chapter 5 is for use in analytical or functional variation modeling (i.e. for cases where significant expertise in the form of hypothesized physical or semi-empirical models or practical expe-
Several key research contributions can be identified in Chapter 4:

- A method of modeling the wafer-level variation compactly using a regression procedure.
- An ANOVA (Analysis of Variance) approach for modeling the die-level and wafer-die interaction terms as a function of the designed layout or process factors. This technique has the added advantage of permitting the determination of confidence intervals on the model coefficients which is not possible using the estimators of Chapter 2.

It is difficult to describe the general methodology of functional or analytical modeling without the use of an example; thus, the components and methods of an analytical variation modeling technique (i.e. test mask development, short-flow processing, and analysis procedures) are offered as a case study in ILD thickness variation modeling in a CMP process in Chapter 5. In the case study,

- Novel test masks and short-flow process flows for characterizing ILD thickness variation in oxide CMP processes are presented.
- The use and application of a semi-physical model towards modeling ILD thickness variation as a function of density for different consumable and process settings is developed.
- The analysis techniques also reveal that pitch, structure area, and perimeter/area do not play a large role in modulating ILD thickness variation in oxide CMP processes.

In Chapter 6, a methodology for determining the impact of variation on circuit performance is demonstrated on three case studies. The methodology is the first attempt at determining the role of pattern-dependent variation on circuit performance. This methodology is enabled via the variation modeling techniques discussed in Chapters 4 and 5. Previous methodologies only considered random normally distributed variation. The methodology is exercised on two case studies on the role of pattern dependent interconnect variation on circuit performance and one case study on the role of device variation on circuit performance. Key contributions of this methodology include:
• Using aerial imaging and other functional models to simulate metal or poly-CD (critical dimension) variation.

• The application of the density model presented in Chapter 5 towards predicting ILD thickness variation.

• Novel point tools which enable the methodology. These point tools include interconnect meshers and gridders, windowed 3-D solver techniques for handling the complexity of large nets, and a “halo” or net-environment extraction procedure for estimating the capacitance of a net to other neighboring nets or structures.

Chapter 7 discusses an example of a variation reduction scheme. In this chapter, the use of metal-fill patterning for improving ILD thickness uniformity is chosen as the case study. Metal-fill patterning is the process of adding metal regions which are not electrically part of the circuit to reduce the changes in underlying topography presented to the CMP tool, and as a result, improve uniformity. Also, a case study is presented which shows a simulated improvement in uniformity of 84% compared to no metal-fill. The material in Chapter 7 is different from other discussions of metal-fill in the literature; here the new research contributions are:

• A unified approach for maximizing the uniformity while simultaneously minimizing any increase in capacitance is used.

• Analysis of the pros and cons of using floating metal-fill (i.e. leaving metal-fill regions unconnected) compared to grounded metal-fill (i.e. connecting metal-fill regions to a fixed voltage source) is offered.

• Two experiments were used to determine the gains possible in uniformity using a relaxed metal-fill patterning rule compared to a very aggressive metal-fill patterning rule.

Finally, Chapter 8 offers some concluding remarks and directions for further research. As represented by the results in each chapter, this thesis is an application of statistical metrology with contributions in each of the three elements of statistical metrology as outlined above.
Chapter 2

Variation Assessment

2.1 Introduction
Relatively little work has been done on spatial variation modeling in VLSI fabrication, with the bulk of the literature focusing on models based on correlation structure or overall wafer-scale variation. Kibarian et al. [9] examined the spatial correlations of device and process parameters (such as threshold voltage, polysilicon line width, and film thicknesses) with circuit performance. Kibarian also proposed methods for extracting wafer level models of variation for these parameters using multiple linear regression techniques. Guo and Sachs [10], and Mozumder and Lowenstein [11] described multi-site response surface methods for modeling within-wafer uniformity as a function of process dependencies. More recently, Davis et al. [12] examined sampling issues for the accurate regression modeling of wafer-scale variation, including methods to utilize multi-site correlation as well as mean-value information. Michael [13] considered the impact of correlated device parameters on circuit performance using simple correlated Gaussian models. Zhang and Styblinski [14] have written extensively on similar correlation models and statistical macromodel development using design of experiment and Taguchi techniques. Yu et al. [7] focused on the details of layout pattern or feature scale variation in polysilicon critical dimensions using a Fast Fourier Transform (FFT) extraction approach. Most recently, Carlen and Mastrangelo have reported on a methodology for characterizing spatial variation in thin film process [15] and for estimating the effect of spatial variation on yield [16].

The methods presented in this chapter are related to the previous work in the following ways. First, we are most interested in spatial variation, and in particular in both wafer-level and die- (or pattern-) level contributions. Second, we are concerned with variance
decomposition rather than establishing correlation between in-line and post-process spatial patterns, or between spatial patterns and process or other dependencies. That is to say, the proposed methods identify and model the specific spatial and die-pattern contributions to parameter variation across the die and wafer. These models can then be used in further correlation studies (e.g. to relate in-line to end-of-line measurements). Finally, the methods we propose make few assumptions about the distribution of the measurements; specifically, we do not assume multivariate correlated normal distributions.

In this Chapter, methods for factoring variation into four main components as defined in Chapter 1: wafer-level variation, die-level variation, wafer-die interactions, and residuals which are assumed to be orthogonal to the systematic components are discussed. Wafer-level variation (the value of some parameter across the surface of the wafer) is often characterized by low (spatial) frequency trends caused by equipment design or operation limitations and is assumed to be independent of layout. Extraction methods for wafer-level variation are presented in Section 2.3. Die-level (or intra-die) variation captures pattern or layout induced deviation of device or structure parameters from their designed values. Methods for die-level variation modeling are detailed in Section 2.4. Wafer-die interaction terms capture differences in intra-die variation as one moves across a wafer. For instance, die near the edge of the wafer tend to have quite different variation profiles compared to die near the center of the wafer. New methods for capturing these interaction terms are introduced in Section 2.5. Methods for analysis of the residuals remaining after systematic components have been removed are presented in Section 2.6. This section further provides a means for the comparison and evaluation of the effectiveness of the decomposition algorithms in Sections 2.3 through 2.5. In addition to presenting the methods used to factor variation, we will also demonstrate the methodology on two datasets. The first is an artificial dataset created to test the efficacy of each of the estimators as they are presented in
Sections 2.3 through 2.5. A description of this dataset is provided in Appendix A. The second example uses data collected from an experiment designed to investigate interlevel dielectric thickness (ILD) variation in chemical-mechanical polishing (CMP) processes. This dataset and analysis results are described in Section 2.7. Finally, concluding remarks are provided in Section 2.8.

2.2 Variation Classification And Definition

Because the physical sources of spatial variation at the wafer- and die-levels are very different, it is critical that methods be available for the separation and analysis of variation components. Equipment and process-related issues can then be identified and addressed via process optimization and control, and pattern dependencies can be minimized by judicious circuit design practices.

Figure 2.1 shows a flow diagram for the general decomposition algorithm we have developed. A hierarchical model is assumed in which the residuals (the output of the previous estimator minus its input) from one estimator become the input to the next estimator. There are three main estimators depicted in Figure 2.1: the wafer-level estimator, the die-level estimator, and the wafer-die interaction term estimator. Detailed descriptions of these estimators are presented in Sections 2.3, 2.4, and 2.5, respectively. The final box in Figure 2.1 represents the residual terms -- the portion of the variation that is left over and assumed to be purely random in nature.
Generally speaking, the variation decomposition algorithm can be expressed in the framework of an additive model. An excellent discussion of generalized additive models, of which we use a special case, can be found in [18, 19]. Using an additive model allows the parameter of interest to be expressed as the sum of several contributions, each with their own distributions or dependencies, such as die-level components \(f_{DLV}\), wafer-level components \(f_{WLV}\), and die-cross wafer level components \(f_{WLV \odot DLV}\):

\[
f_{RAW} = f_{WLV}(x, y) + f_{DLV}(x, y) + f_{WLV \odot DLV}(x, y) + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma^2) \tag{2.1}
\]
In (2.1), $x$ and $y$ are spatial coordinates on the wafer and $\varepsilon$ corresponds to the residual terms mentioned before.

2.3 Wafer-level Variation Extraction

The wafer-level variation is assumed to originate from large-scale equipment and process asymmetries. These may result from the equipment design (e.g., asymmetric chamber geometry affecting gas flows, thermal gradients in a furnace) or from equipment/process imperfections (e.g. nonuniform slurry transport in chemical mechanical polishing). Because of the underlying mechanisms, the wafer-level variation is assumed to be slowly varying and smooth. Simple methods for extracting the wafer-level variation are often employed, including examination of only a single identical structure on every die, or use of a simple moving average on all of the available data. Unfortunately, these simple methods are often ineffective. Examination of only one structure on each die neglects the large amount of information contained in other structures within the die. Also, this measure implicitly assumes that the physical mechanism behind wafer-level variation is not associated with spatial distances less than the die size as might occur in cases of severe wafer edge effects. Finally, a simple moving average can yield poor estimates near the edge of the wafer where the window of the moving average is not completely located inside the wafer, necessitating careful corrections. For these reasons, alternative methods are often needed to extract the wafer-level variation.

In this section, we evaluate, analyze, and recommend four methods for estimating the wafer-level variation: the downsampled moving average estimator (DSMA), the meshed spline method (MSM), linear regression coupled with a novel physically based cross validation approach, and an estimator based on the linear combination of these techniques. After describing the background behind each of these methods, the strengths and weak-
nesses of each method is explored, and an artificial dataset is used to illustrate and gauge the relative utility of each wafer-level estimator.

2.3.1 Downsampling Moving Average Estimator (DSMA)

The downsampling moving average estimator (DSMA) utilizes the intuitive notion of a moving average to smooth over rapidly varying features arising due to die effects. This procedure (and all those considered in this Chapter) begins by taking the raw dataset and interpolating it onto a regularly spaced grid. The size of the grid is chosen based on the number and distribution of observations in each die in order to avoid grossly quantizing the data; e.g. approximately \( n \) grid lines per die are chosen in each direction for \( n^2 \) data points scattered throughout a die. The interpolated input grid is then downsamped to generate a lower density grid (e.g. 4x4 downsampling to create a grid composed of only every fourth grid point in the \( x \) and \( y \) directions compared to the original grid). Values for the downsamped grid points are then computed by taking an average of the full dense grid data within a prescribed distance of the corresponding original grid point; the neighborhood used is usually the same as the degree of downsampling. These averaged downsamped data points are then used to interpolate back to the original size of the input dataset. Effectively, this interpolation step is similar to upsampling: it replaces all of the “missing” data points of the original dataset removed during the downsampling procedure with estimates calculated via interpolation. The recommended interpolation scheme is a cubic fit to each surface formed by the triangulation of data-points and partial derivative information [20]. Finally, for each data point, the difference between the raw data and its corresponding interpolated value is computed and regressed onto a function of \( x, y \), and \( xy \) to capture any remaining systematic linear trends in the dataset. The result of this regression is added to the aforementioned interpolated values to form the DSMA estimate.
The effect of the DSMA estimator procedure is best understood through examination in both the discrete space and frequency domains. Figure 2.2 illustrates key properties of the DSMA estimator for a one dimensional example. Figure 2.2a and 2.2b show the input sequence (left side) and the magnitude of the Fourier transform of the input sequence (right side). The spikes in Figure 2.2b are caused by the die periodicity of the input sequence. The DSMA estimator can be thought of as the convolution of a moving average (Figure 2.2c) with the input sequence followed by a downsampler (Figure 2.2e). The Fourier transform of the moving average is a sinc function, where the moving average window size ($M$ in Figure 2.2c) determines the location of the zeros in the frequency domain. Thus, if the size of the window is chosen to be an integer multiple of the die period, we see that the spikes caused by the periodicity in the input sequence will be zeroed out. Note, the linear regression step discussed in the DSMA estimation procedure is not shown in Figure 2.2. This benefits of the step are most readily seen for two dimensional examples.

In the spatial domain, downsampling of order $W$ can be represented by multiplication with a periodic impulse train of period $W$, while in the frequency domain, downsampling corresponds to a convolution of the output of the moving average with a periodic impulse train of period $N/W$ where $N$ is the length of the input sequence. This convolution step causes the main lobe shown in Figure 2.2b to be accentuated and replicated at each impulse while other components are either attenuated or aliased. The final interpolation step replaces the missing data points in Figure 2.2g with estimates derived via interpolation. In the frequency domain, the interpolation step appears as the multiplication between Figure 2.2f and the impulse response of the interpolator (essentially a low pass filter) so that only one slightly filtered and attenuated copy will survive. This cubic interpolation is an important step often omitted in simple die-average approaches, and serves to extend the gentle wafer trends better near the edge of wafer.
In the DSMA estimator, the amount of aliasing must be controlled to avoid a poor estimate of the wafer-level variation. Aliasing occurs when the spikes due to the periodic nature of the input (Figure 2.2b) are not sharp but appear as rounded lobes and are con-
volved with the sampling step (Figure 2.2f). This condition occurs in cases of large wafer-die interaction terms or pseudo-periodicity in the input sequence. In those cases, a different estimator (spline or regression fit) should be chosen.

The downsampling rate is another critical factor. If the downsampling rate is made too fine (less downsampling), less aliasing will occur since the period of the impulse train in Figure 2.2f will increase, but problems will occur at the edges of the estimator due to the interaction between the moving average and the finite length of the input data sequence. That is, the estimator will react too strongly to a small number of data points near the edge of the wafer. We propose guidelines for choosing the downsampling rate in Section 2.3.6.

2.3.2 Meshed Spline Method (MSM)
The meshed spline method (MSM) is distinguished from the DSMA primarily in that non-parametric estimates of values on selected spatial points along lines of interest are generated, rather than point-wise averaging and filtering. In the MSM, the input data is again first interpolated onto a regular grid with the grid size determined by the number of observations per die. The procedure then selects grid lines or “stripes” (every $M^{th}$ line), and smooths the data points for that stripe using spline methods. This sampling scheme is essentially the same as the downsampling in DSMA, except in the MSM we downsample in one direction at a time (by picking each $M^{th}$ grid line) and spline in the other direction (along that grid line). Spline smoothing is a non-parametric based technique for fitting data and is summarized in Appendix B. In our application, the spline smoothing procedure is done first on horizontal stripes followed by vertical stripes. The results of these two steps are then averaged together to eliminate ambiguity associated with points that fall on both horizontal and vertical stripes. After the stripe and smooth procedures, the estimate is passed through the same cubic interpolation and linear regression procedures used for the DSMA estimator. The MSM can also be repeated for meshes oriented in other orthogonal
directions such as along 45 and 135 degree lines.

Figure 2.3 illustrates the key steps involved in the MSM estimator; again the spatial domain is shown on the left and the frequency domain on the right. Artificial raw two-dimensional data is pictured in Figure 2.3a, with the corresponding 2D Fourier transform in Figure 2.3b. The mesh and smooth operation for horizontally oriented lines can be written as:

**Figure 2.3:** Meshed spline method (a, b) the input sequence, (c, d), the result of striping in the y direction, (e, f), the result of striping in the x direction, (g, h) the MSM estimate for the wafer-level variation shown in (a). (b, d, f, h) is the frequency response of (a, c, e, g)
where \( h_{int}(x,y) \) is the impulse response of the spline interpolator\(^1\), \( \otimes \) represents the convolution operator, and \( \delta[x] \) is the unit impulse function:

\[
\delta[x] = \begin{cases} 
1 & \text{if } x = 0 \\
0 & \text{if } x \neq 0
\end{cases}
\] (2.3)

The 2-D Fourier transform of (2) can be written as:

\[
F_{WLV}[k_x, k_y] = \sum_l \left( H_{int}[k_x, k_y] \cdot F_{RAW}[k_x, k_y] \right) \otimes \left( \delta[k_x - iN/M] \cdot \delta[k_y] \right). \tag{2.4}
\]

Thus, the periodic nature of the peaks shown in Figure 2.3d and 2.3f is due to the sampling process inherent in the mesh and smooth operation while the higher-order sinc-like behavior is due to the interpolative action of the mesh and smooth operation. When Figure 2.3c and 2.3e are averaged together and interpolated to form a smooth surface in the spatial domain, only the center peak (highlighted in Figure 2.3d and 2.3f) is retained in the frequency domain.

### 2.3.3 Regression Based Estimators

For regression estimators, the wafer-level variation is assumed to be modeled by a smooth polynomial functional of \( x, y, r, \theta \), or other relevant coordinates, e.g.

\[
f_{WLV} = ar + br^2 + c\theta \tag{2.5}\]

where standard least squares regression to the available data may be employed. Mixing of coordinate terms \((x, y)\) and \((r, \theta)\) should be avoided to prevent collinearity and structured variance-covariance matrices.

---

1. Technically speaking, a spline interpolator is a non-parametric technique, and thus, it does not have an impulse response nor is it a linear operator. To a first approximation, any interpolator can be represented as an impulse response with a sinc-like functional form. Further research, however, is needed to fully explain the behavior of the MSM estimator.
In order to determine the best model for the data, a unique method of process experimentation coupled with standard cross-validation techniques is proposed which we term “physical cross-validation.” If multiple wafers with the same pattern are being analyzed, standard cross-validation approaches (e.g. leave one out or other data subsetting) together with standard model selection procedures (e.g. forward, backward, stepwise, and all subsets) may be employed to construct the wafer level model. However, one must be careful in assuming that the resulting wafer-level variation is truly independent of the pattern: the wafer “as a whole” may or may not look the same to the process with and without patterns (or with different patterns) on it. Thus, additional physical validation on wafers with different patterns or on unpatterned (or “blank”) wafers is suggested. Unpatterned wafer analysis is also valuable in confirming that the wafer-level model is not capturing die-level interaction components, as both die-level and wafer-die interactions are not present. In the case of inter-level dielectric thickness modeling, the blank wafer approach might constitute depositing a new wafer with SiO\textsubscript{2} and polishing with CMP, or in the case of photolithography modeling, blank wafers could be coated with photoresist and measurements of photoresist thickness across the wafer might be collected. After processing, standard cross-validation and model selection approaches can again be employed to construct and check a representative wafer-level model. In any multiple wafer approach, one must also account for or block against wafer-to-wafer variation due to process/equipment drift in time.

2.3.4 Combined Techniques

A linear combination of the estimates presented in the previous sections may also be used to determine the wafer-level variation. Thus, if each wafer-level estimate presented in this section is labeled \( E_1, E_2, \) and \( E_3 \) then a combined wafer-level estimate, \( f_{WLV,eff} \), can be formulated as:
\[ \epsilon_{\text{WLV, eff}} = w_1E_1 + w_2E_2 + w_3E_3 \]  

where \( w_1, w_2, \) and \( w_3 \) are weights (normalized to sum to one). Without other \textit{a-priori} knowledge, the weights can be set equal, \( w_1 = w_2 = w_3 = 1/3 \). If, however, it is known that one estimator is more likely to yield a more efficient\(^1\) estimator compared to the others, then it can be more heavily weighted. Assuming that the errors associated with each wafer-level estimator are independent and normally distributed with roughly the same variance, the combined estimator should yield an estimate with a smaller variance by approximately a factor of \( (w_1^2 + w_2^2 + w_3^2) \).

---

1. In this work, the efficiency of an estimator is defined as: 
\[ \text{Efficiency} = \frac{\text{var}(\text{Estimate} - \text{Actual})}{\text{var}(\text{Estimate})} \]
2.3.5 Simulation Results

An artificial dataset (shown in Figure 2.4) is used to evaluate the estimators presented in this chapter; the details of the dataset can be found in Appendix A. The results of extracting the wafer-level variation for each of the estimators described in the previous sections are shown in Table 2.1. In this Table (and elsewhere in this Chapter), the “error standard deviation” is the standard deviation of the difference between the actual and the extracted value, and the “% error” is defined as the ratio of this error standard deviation to the mean.

*Figure 2.4:* Artificial data set (described in detail in Appendix A)
of the raw value, times 100. The DSMA estimator appears to be the best with respect to these measures, followed by the MSM, with the regression based estimator the least effective. The results shown in Table 2.1 require careful interpretation, however, as it was empirically found that the relative efficiency of the MSM estimator versus the DSMA was a function of the magnitude and distribution of the wafer-die interaction terms. This phenomenon can be attributed to aliasing effects in the DSMA estimator. In the MSM estimator, less aliasing occurs since the spline smoother effectively eliminates most of the high frequency components which could lead to aliasing effects during any resulting upsampling steps. For comparison, also shown in Table 2.1 is the result from a simple wafer level estimator, termed “single structure method”, in which only a single structure (the same structure) on each die is measured and the cubic interpolator is used to obtain wafer-level values for all other locations on the wafer.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Standard Deviation</th>
<th>%Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downsampled Moving Average</td>
<td>0.099</td>
<td>1.4%</td>
</tr>
<tr>
<td>Multiple Spline Method</td>
<td>0.363</td>
<td>5%</td>
</tr>
<tr>
<td>Regression Method</td>
<td>0.546</td>
<td>6.4%</td>
</tr>
<tr>
<td>Single-Structure Method</td>
<td>0.520</td>
<td>8%</td>
</tr>
</tbody>
</table>

2.3.6 Parameter Selection Guidelines for the DSMA and MSM Estimators

There is some ambiguity associated with the amount of downsampling required in the DSMA and MSM estimators. We propose to resolve this ambiguity by a new method in which we estimate a “characteristic wafer-level variation length” based on the average correlation distance. (Alternative values for the wafer-level variation length could be based
on several blank wafers which have been placed through the processing equipment that one is trying to model. The data from these wafers are collected and the average correlation distance is computed based on just this wafer-level variation. The average correlation distance is computed by obtaining the one dimensional discrete autocorrelation function (ACF) of several slices through the wafer diameter chosen randomly. The correlation distance is found for each ACF slice by finding the lag (a discrete unit based on the uniformly spaced grid) corresponding to where the ACF has fallen to the 95% confidence level about zero [21]. These correlation distances for each slice are then averaged together to form the characteristic wafer-level variation length, $L_{WLV}$. (If the correlation “distances” are measured in discrete units, they can be converted to the continuous spatial dimension by multiplication by $\Delta x$ and $\Delta y$, the uniform grid spacings in the $x$- and $y$-directions, respectively.)

$L_{WLV}$ however, might be non-symmetric in that the average correlation distance for $x$-oriented slices through the wafer might be significantly different than the average correlation distance for the $y$-oriented slices. In this case, two characteristic lengths should be defined: $L_{WLV,X}$ for the average correlation distance in the $x$-direction and $L_{WLV,Y}$ for the $y$-direction. Correlation distances indicate the amount of smoothness that will be required to reject unwanted noise or to reject data which is not associated with the underlying physical phenomenon causing wafer-level variation. A good choice for the downsample rate in the DSMA and MSM estimators is the simultaneous use of these two correlation dis-
stances, or \([L_{WLV,X} \times L_{WLV,Y}]\); in this way, each sample will contain or represent data which is well correlated with that sample, but not include data beyond that correlation distance.

Figure 2.5 shows the discrete ACF for a particular slice through the artificial wafer-level variation dataset with a resulting computed correlation distance of approximately 7 or 8. Other slices yielded similar results resulting in \(L_{WLV,X} = L_{WLV,Y} = 8\). As Table 2.2 shows for the DSMA estimator, decreasing the degree of down sampling initially leads to a better estimate of the wafer-level variation as measured by the error standard deviation; after a critical threshold is reached, however, decreasing the down sampling further leads to excessive leakage of die-level energy into the wafer-level component and hence a poor estimate. This empirically observed optimal choice (see Table 2.2) of 6x6 downsampling corresponds well to the value of 8x8 calculated from the characteristic wav er-level variation lengths.

Figure 2.5: Autocorrelation function for artificial data set
2.3.7 Discussion of Wafer-level Variation Extraction Techniques

The primary advantage of the DSMA estimator is simplicity and efficiency. If only one or two sites per die are available, it is still possible to compute the wafer-level estimator. The main disadvantage of this technique is that rapid but systematic curvature changes at the edges of the wafer are likely to be missed if relatively few points per die are used in the estimation process. In many processes, however, rapid curvature is not expected, and in such cases the DSMA estimator is generally the most efficient estimator for extracting wafer-level variation.

The primary advantage of the MSM estimator is that rapid edge effects can be more readily captured. The main disadvantage of this procedure is that it can be computationally expensive: the number of points which must be fit via the spline is generally about an order of magnitude greater than the number of points used in average computation for the DSMA. In addition, spline smoothing is known to result in some anomalous behavior if the degrees of freedom parameter (see Appendix B) is set improperly.

The regression technique holds one significant advantage compared to the previous wafer-level extraction methods. Regression produces a parametric estimator so that the

<table>
<thead>
<tr>
<th>Downsampling Rate</th>
<th>Error Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16x16</td>
<td>0.447</td>
</tr>
<tr>
<td>14x14</td>
<td>0.240</td>
</tr>
<tr>
<td>12x12</td>
<td>0.121</td>
</tr>
<tr>
<td>10x10</td>
<td>0.137</td>
</tr>
<tr>
<td>8x8</td>
<td>0.128</td>
</tr>
<tr>
<td>6x6</td>
<td>0.100</td>
</tr>
<tr>
<td>4x4</td>
<td>0.208</td>
</tr>
<tr>
<td>2x2</td>
<td>0.407</td>
</tr>
</tbody>
</table>

Table 2.2: Effect of Downsampling on DSMA Estimator Accuracy
resulting model is very compact and the model coefficients for different wafers and lots are useful tools for tracking lot-to-lot and wafer-to-wafer variation. The primary drawback with regression techniques is that they assume a large degree of symmetry or analytic regularity in the wafer-level variation which is often not the case. The other techniques presented in this section are more robust since they are based on non-parametric methods.

2.4 Die-level Variation

This section focuses on die-level variation extraction. The estimation method here draws upon original contributions by Yu and Spanos in [15]; this frequency domain based approach is first reviewed. The results of applying the die-level estimator to our artificial dataset are then presented. Finally, the importance of spectral interpolation issues not discussed in detail in [15] is examined: the purpose and role of the spectral interpolation step is detailed, and guidelines are developed for the minimum number of die on a wafer needed to extract an efficient die-level variation estimator.

2.4.1 FFT Based Method

The FFT-based die-level estimator is fundamentally based on detecting the periodicity inherent in the spatial repetition of die on a wafer. The residuals from the wafer-level variation, \( f_{RAW} - f_{WLV} \) in the additive model presented in (1), are first transformed via the two dimensional fast Fourier transform (FFT). The Fourier transformed data is then filtered (passing components near the die frequency) to capture the components due to die-level variation sources. These captured components are placed into a separate matrix and the inverse Fourier transform is computed to produce the extracted die-level variation.

The Fourier transform of the input data stream \( f[x, y] \) is \( F[k_x, k_y] \) where
Since intra-die variation is mostly periodic with a spatial period of the die size, the majority of the die-level variation terms are assumed to be concentrated at $N_x/D_{P_x}$ and $N_y/D_{P_y}$ multiples of the sampling period where $N_x, N_y, D_{P_x},$ and $D_{P_y}$ are the number of samples in the $x$- and $y$-directions, and the die-period in the $x$- and $y$-directions, respectively. These components form the basis for the die-level estimate.

A key issue with the FFT die-level estimator is how many die are needed to obtain an accurate estimate of the die-level variation. This question can be answered by considering $x[n] = m[n] + e[n]$ where $e[n] \sim \text{N}(0, \sigma^2)$ and $m[n]$ is a periodic signal of length $N$. Thus, $X[k] = M[k] + E[k]$ where $X[k]$, $M[k]$, and $E[k]$ is the Fourier transform of $x[n]$, $m[n]$, and $e[n]$. If we have $M$ replicates of a signal $x[n]$ each of which is of length $N$, it can be shown that the $\text{var}(X[k]) = \text{var}(M[k]) + \sigma^2 MN$ [22]. Consider $\text{var}(X[0])$ (which is usually the largest Fourier term assuming the mean has not been removed) $= (\mu MN)^2 + \sigma^2 MN$. In this way, a constraint on the signal to noise ratio for $X[0]$ can be expressed as

$$\frac{\sigma \sqrt{MN}}{\mu MN} < \alpha \quad (2.8)$$

We can express the required number of replicas of a signal (i.e. the number of “die”) $M$ as:

$$M > \left( \frac{\sigma}{\alpha \mu} \right)^2 \times \frac{1}{N} \quad (2.9)$$

Figure 2.6 shows an example used to verify (2.8). In this example, $\sigma = 0.1$, $\mu = 1/(2N)$, and $M = 100$ resulting in an $\alpha$ of 0.057 compared to a theoretical value of 0.063.

In terms of the die-level estimator, $\sigma$ can be estimated by taking the site variance (i.e. the variance of all measurements from the same location on the die) and $\mu$ can be esti-
mated from the mean of the raw variation (assuming that the data has not been detrended).

Typical targets for $\alpha$ are 1%. Considering two dimensional effects with this terminology leads to the design constraint,

$$\frac{M}{\sigma} > \frac{1}{\sqrt{N}}$$  \hspace{1cm} (2.10)

In [15], an interpolation scheme was suggested to prevent the wafer-level components from aliasing onto die-level estimates. Since the wafer-level components have already

\textbf{Figure 2.6}: (a,b) A sample signal and Fourier transform used to evaluate the accuracy of (2.8). (c) Estimates of $X[0]$ drawn from 100 independent trials.
been removed, there is little need to interpolate and interpolation actually does more harm than good when the majority of the wafer-level components have been removed.

2.4.2 Simulated Dataset

In our artificial dataset (shown in Figure 2.4a), the repetitive structure is due to the die-level variation. A detailed view of the known artificial pattern for one die is shown in Figure 2.4c; since the imposed die-level variation is periodic all other die are identical to Figure 2.4c. Table 2.3 shows that the extracted die-level variation is nearly identical to the actual die-level variation in Figure 2.4c. Similar results have been obtained for other artificial datasets in which the die-level variation is known.

Table 2.3: Model Dataset: Die Level Variation Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>Error Standard Deviation</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLV\text{extracted} - DLV\text{actual}</td>
<td>0.045</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

In the die-level variation estimator presented in this section, periodicity of intra-die variation components is always assumed. For short-flow processes, this is a good assumption since the confounding of intra-die variation is localized to a relatively small number of processing steps. In a full-flow process, however, periodicity is often not observed due to the repeated confounding and small offsets from many processing steps. The die-level variation estimator presented here is also hampered by the constraint, required for the FFT computation, that large numbers of die must be probed in a strictly periodic fashion. Methods for circumventing this restriction are presented in Chapter 4.

2.5 Wafer-die Interaction Terms

Two methods are presented below for estimating the wafer-die interaction terms. The first is a simple spline based method, while the second is a frequency domain based proce-
dure utilizing further FFT analysis. After summarizing the two methods, the relative efficiency of each estimator based on our artificial dataset is presented. In each case, the input to the estimator is the raw data minus the extracted die-level and wafer-level components, or in the framework of the additive model presented in (1), $f_{\text{RAW}} - f_{\text{WLV}} - f_{\text{DLV}}$.

2.5.1 Spline Based Method

In the spline based method, the die-level residual terms (those remaining after the wafer-level and periodic die-level components have been removed) are spline smoothed in an effort to distill out any remaining systematic variation. The uniform grid is processed one row at a time with the parameter degrees of freedom (see Appendix B) controlling the degree of smoothness or roughness. Figure 2.7 shows the result of smoothing a one dimensional slice from a real dataset for several choices of degrees of freedom. A small value for the degrees of freedom (e.g. 5 in Figure 2.7) leads to an excessively smooth and over-filtered result. We define this condition as undersaturation. A large value for the degrees of freedom (e.g. 200 in Figure 2.7) leads to excessive noise leakage and limited smoothing which also limits the effectiveness of the procedure. We call this condition oversaturation.
Choosing the proper degrees of freedom is important, but the optimal choice of degrees of freedom is difficult to define precisely (see Appendix B). The exponential spar is one approach for obtaining a reasonable degrees of freedom parameter. In this approach, a representative one-dimensional slice of the data is smoothed for all choices of degrees of freedom from 2 to $N-1$ where $N$ is the discrete length of the slice. For each choice of degrees of freedom, the spar value $\lambda$ is computed (see Appendix B) [21, 24]. A plot of $\ln(\lambda)$ versus degrees of freedom is then generated, from which the final degrees of freedom is selected by finding the tangent point of this graph to a 45 degree line (Figure 2.8). The interpretation of this procedure is that a significant decrease in the degrees of freedom leads to excessive smoothness while a significant increase in the degrees of freedoms lead
to rough and noisy surfaces. While some arbitrariness remains, good results have been generated using this method.

![Figure 2.8: Exponential spar method](image)

The primary weakness of the spline based method is the present restriction of the smoothing to one dimension only. Also, the lack of a precise determination of the optimal choice of degrees of freedom beyond the accuracy permitted by the simple exponential spar method leads to additional errors. Some improvement can be obtaining using the striping methods developed in the wafer-level MSM, but better results might be possible using a two dimensional model fitting method such as loess regression [25], generalized linear modeling [21, 26], or kriging techniques [27, 28, 29].

### 2.5.2 FFT Based Method

A second approach to identifying die-wafer interactions in the variation is based on an extension of the Fourier transform methods in Section 2.4. After transforming the die-
level residual data to the frequency domain using an FFT, a subset of the spatial frequency components are selected (allowed to pass into the interaction estimator). These points are then transformed back into the spatial domain using an inverse Fourier transform. A key issue associated with this estimator is choosing which frequency components to select. The components to select depends greatly on the raw variation source. If it is know or believed (and verified by inspection) that the variation source is spatially stationary\(^1\) (i.e. wafer-to-wafer and lot-to-lot variation only has a strong temporal component and very little spatial dependence) and several wafers have been measured, a good practice is to compute the pointwise average for every measured wafer (i.e. to average all measurement sites across all wafers and lots) and perform the decomposition on this averaged wafer. Since the data is assumed spatially stationary, the residual components for this “averaged” wafer will be averaged out and the die-level residuals (i.e. the raw variation minus the wafer-level variation minus the die-level variation) as a whole form an estimate for the interaction terms. The residual components for a particular wafer is then determined by computing the die-level residuals for a particular wafer of interest and subtracting the interaction term estimate (that was extracted from the “averaged” wafer) to form the residual.

If the variation source is spatially non-stationary, there are two options. The first technique involves computing the die-level residuals for each measured wafer. The interaction term is then formed by averaging the die-level residuals for each measured wafer across all wafers. If there are not many measured wafers, this technique will not be effective since the averaging power of the estimator goes as the inverse of the square root of the number

---

1. A signal is spatially strict sense stationary if all the moments of the signal do not depend on the position on the wafer. This is to say that the variability of a site does not depend on position. Clearly this assumption is violated if there is a large wafer-to-wafer or lot-to-lot variation. A signal is wide-sense spatially stationary if only the mean and the variance are invariant of site location. For the estimator discussed here, only strict sense stationary is needed if the underlying site variation or random noise is known to be white or normally distributed. Non-normal or shaped noise or residuals require strict sense stationary.
of wafers. In these cases, the only remaining technique is to take the sidelobe components of the die-level residuals. These are the components which surround the frequency components which were extracted using the die-level estimator. The rational behind this technique is that probabilistically speaking these terms are very likely to contain systematic energy especially when the “true” interaction terms have a pseudo-periodic component.

Figure 2.9 shows two cases which illustrate these methods. Figure 2.9a shows the die-level, wafer-level, wafer-die interaction, residual, and raw components of a simple “one-dimensional” wafer. The results generalize for two-dimensional data, but only a one-dimensional example is shown for clarity. The “true” wafer-die interaction terms were formed by modulating the die-level component with the wafer-level component and scaling the result. The residual terms are treated as a white-noise source.

Figure 2.9b shows the frequency spectrum of the die-level residuals and the spectrum of the “true” interaction terms for a case where the residual term is small. The two spectrums are nearly identical except for some visible noise and the obvious discrepancy at the fundamental frequencies of the die period. This discrepancy is due to the die-level estimator which removes these frequency components during estimation. The “true” interaction terms are pseudo-periodic with the die period; thus, we expect to see some non-zero energy at the harmonics of the die-period. Lumping these terms into the die-level components is not problematic since the interaction terms contain a periodic component which is modulated by a non-periodic component to result in pseudo-periodicity. Thus, the perfectly periodic components of the data are lumped with the die-level variation and any systematic non-periodic components which are not in extracted as part of the wafer-level variation are lumped into the interaction terms.
**Figure 2.9:** (a) The wafer-level, die-level, interaction, residual, and raw variation components for an illustrative one-dimensional wafer. (b) the frequency spectra of the known interaction terms versus the die-level (or averaged die-level) residuals for (b) a small residual component, (c) a large residual component, and (d) a large residual component but averaged across many wafers.
For the example shown in Figure 2.9b, the interaction terms can be estimated by lumping the entire die-level residuals as the interaction terms. When the residual components are much larger, however, as shown in Figure 2.9c, the frequency spectrum of the die-level residuals and known interaction terms differ significantly. Figure 2.9d shows the result of averaging together the die-level residuals for several one-dimension wafers. The result is now much closer to the known interaction term spectrum. The residual terms are then estimated by subtracting this interaction term estimator from the die-level residuals for the particular wafer of interest.

The main advantage of this FFT method is robustness. Unlike the spline based method, no parameters such as degrees of freedom need to be adjusted and tuned. The main source of error of this method comes from the inability of the FFT to reject all of the noise terms even with spatial averaging. In addition, this procedure is limited by the same restrictions on the FFT die-level estimator presented in Section 2.4.

2.5.3 Simulated Extraction and Error Analysis

Figure 2.10 shows a one dimensional slice through an artificial dataset (not related to the dataset presented in Appendix A). Figure 2.10a shows the input to the estimator, and Figure 2.10b shows the extracted wafer-die interaction terms. The error term and the autocorrelation function for the error are shown in Figure 2.10c and 2.10f respectively. The degrees of freedom parameter was determined from Figure 2.10d using the exponential spar method, and set to 45. As the residual autocorrelation function (ACF) shows, the estimator has done a good job in removing most of the systematic sources of variation that are indicated by the ACF (Figure 2.10e) for the original input data (Figure 2.10a).
In contrast to the simulated dataset in Figure 2.10 which has a relatively small wafer-die interaction term compared to the wafer- and die-level terms, our artificial dataset of Figure 2.4 incorporates a large wafer-die interaction term. As Table 2.4 shows, both wafer-die interaction term estimators exhibit similar performance characteristics.

**Table 2.4: Simulated Wafer-Die Interaction Term**

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Standard Deviation</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spline Method</td>
<td>0.0953</td>
<td>1.1%</td>
</tr>
<tr>
<td>FFT Based Method</td>
<td>0.0826</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Figure 2.10: Wafer-die interaction terms for an artificial data set. Extracted using spline based estimator.
2.6 Residual Analysis

In this section, two methods are presented for determining the size and content of any remaining systematic variation in the residual terms. The residual terms are the remaining components after passing through the wafer-die interaction estimator (Figure 2.1). Using the notation of (1), the total residuals are $f_{RAW} - f_{WLV} - f_{DLV} - f_{WLV \odot DLV}$. The correlogram [29, 31] assesses the extent of spatial correlation remaining in the residual terms. The die-mean residual test is a simple procedure for identifying the degree of periodic or pseudo-periodic data still left in the total residual term. These two approaches are discussed in further detail below. In addition to these two analyses, examination of the distributions of the residuals using standard methods such as quantile-quantile plots is also recommended but not detailed further here [32].

2.6.1 Correlogram Methods

The correlogram (and related semi-variograms and variograms) [29, 31] of the residual component is useful for assessing the presence or absence of systematic spatial components. In a correlogram, two-dimensional spatial correlation is plotted versus distance lag. Large positive or negative numbers indicate the degree of spatial correlation for all points located within the specified spatial distance (as read from the spatial lag axes) with +/- 1 (on the spatial correlation axis) indicating perfect correlation and 0 indicating no correlation. In this way, the correlogram of a residual component with no systematic components remaining should cluster around zero at all spatial lags while values which are non-zero indicate significant correlation. The primary drawback with the correlogram is it does not capture the relative magnitude of the residual term compared to any of the other terms. As a result, the correlogram can exhibit relatively high residual correlations even though the total residual term is 1000 times or more smaller than the original raw dataset. Also, direc-
tional variograms and correlograms [29, 31] can be used to identify any anisotropy in the data.

2.6.2 Die Mean Residual Test

The die mean residual test is a statistic useful for determining the strength of any systematic periodic or pseudo-periodic components still present in the total residuals. In this method, the average of the same location on each die is computed for all points in a die. If a die is represented by 24x24 data points with 5x6 die per wafer then 24x24 mean values will be computed, and each mean will be an average of 5x6 or 30 points. Alternatively, the median can be used. This procedure can be written compactly as

\[
\overline{f}(x, y) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} f(x_{m+iT_x}, y_{n+JT_y}),
\]

where \( N_x \) and \( N_y \) are the number of die in the \( x \)- and \( y \)-directions, and \( T_x \) and \( T_y \) are the die-period in the \( x \)- and \( y \)-directions. Using the previous example, \( N_x = 5 \), \( N_y = 6 \), \( T_x = 24 \), and \( T_y = 24 \).
Since averaging is involved in the calculation, any systematic data should be amplified while random noise should be averaged out or attenuated. In this case the magnitude of any residual component is properly represented: small systematic sources of variation lead

1. This result is only true for zero mean Gaussian or other balanced distributions. Also, if there are periodic components with periods greater than the die size, the DMRT will yield invalid results.
to small signals using the die mean residual. The drawback to this method is that unlike the ACF method, no statement about correlations can be made from the die mean statistic. Because of this, a careful application of both analyses is useful. Also, the die mean residual test cannot reveal the source of any remaining systematic variation (i.e. wafer-level vs. die-level).

2.6.3 Simulation Results

Table 2.5 summarizes the statistical properties of the total residuals extracted from our artificial dataset described in Appendix A. Figure 2.11a shows the quantile-quantile plot (qq-plot) [32] for the total residuals. The qq-plot shows that the total residual is roughly Gaussian for data within $\pm 2\sigma$. Note that the extreme volume of data available allows greater examination of the tails than is usually possible. Figure 2.11b-c shows the correlogram and die mean residual test (DMRT) (for one die) for the total residuals extracted from the artificial dataset (Figure 2.4). With the exception of very small spatial lags, the correlogram shows very little spatial correlation. Figure 2.11c shows the DMRT results; the plot is flat indicating that there is relatively little remaining residual systematic variation.

<table>
<thead>
<tr>
<th>Property</th>
<th>Extracted Value</th>
<th>Actual Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual Mean</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Residual Standard Deviation</td>
<td>0.13</td>
<td>0.10</td>
</tr>
</tbody>
</table>
2.7 Experimental Dataset.

Figure 2.12: HP Mask Design Showing (a) die layout (area intensive structure lower right), (b) subdie layout: four capacitors with layout factors near local line-width measurement structures, and (c) capacitive ILD thickness test structures with (d) layout experimental design factors

In this section, the procedures outlined in the previous sections are applied to an experimental dataset from a CMP short-loop test die fabricated at the HP ULSI fabrication
facility in Palo Alto, CA [30]. The experiment is designed to investigate intra-die variation of interlevel dielectric thickness remaining after chemical mechanical polishing (CMP). The test mask is briefly summarized, and the resulting wafer-level, die-level, and wafer-die interaction term estimates are presented, followed by consideration of the remaining residuals.

Figure 2.12 shows the 1.45 cm x 1.45 cm short-loop test die. A spectrum of electric test structures with different combinations of layout dimensions to probe for feature and pattern dependencies populate three of the four die quadrants. The fourth quadrant contains additional test structures with the same feature scale (e.g. line width and space) factors but different areas. These test structures were fabricated on 6” wafers containing fifty-four die. Two lots of twenty wafers each were fabricated. The first lot was fabricated in 1994 using a particular CMP processing tool (Polisher A) while a second lot was fabricated in 1995 using a different CMP processing tool (Polisher B). Three wafers (M2, M5, M6) are considered from the 1994 lot and two wafers (B5, B9) were probed from the 1995 lot. All available structures and die were sampled. While this is not required, empirical results have suggested that the majority of the dies need to be sampled with at least five sites, preferably sampled uniformly across the die, per die sampled. Further details on the experimental methodology can be found in [5].
The procedures and algorithms outlined in Sections 2.3 - 2.6 were implemented in S-PLUS TM [33] along with a simple user interface. The resulting analysis tool, entitled VarDAP, or Variation Decomposition Analysis Program, was used to decompose the observed variation into its constituent parts and produce the following results.

Figure 2.13: Wafer-level variation, die level variation, interaction terms, and total residuals for a die near the center from the HP Mask Set
Figure 2.13 shows the extracted wafer-level, die-level, wafer-die interaction, and residual variation corresponding to one die for wafer B5. Figure 2.14 shows example extracted wafer-die interaction components. For these figures, the DSMA and FFT based methods were used, and the same z-axis scaling is used. Figure 2.14a shows the wafer-die interaction term for one die near the center while Figure 2.14b shows the interaction term for another die near the edge of the wafer. A significant edge effect (clearly larger than residual component) is visible.

A histogram for the raw data, wafer-level variation, die-level variation, wafer-die interactions, and total residuals for wafer B5 is shown in Figure 2.16. We see that the total residual is significantly smaller than the original raw data distribution, indicating that the extraction methods have been successful in identifying key systematic spatial variation.
components. In addition, the total residual is primarily Gaussian over ±1.5σ as evidenced by the quantile-quantile plot (Figure 2.16f) while there is still some remaining non-Gaussian components in the tails of the distribution. The result of applying the die mean residual test on wafer M2 of the experimental dataset is shown in Figure 2.15, and as the magnitude of the output (plotted on the same scales) suggests, the vast majority of the systematic pseudo-periodic and periodic sources of variation have been extracted.

![One Dimensional Slice Through Wafer Diameter](image)

![ILD Thickness](image)

![One Die Perspective Plots](image)

**Figure 2.15:** Die Mean Residual Test for Wafer M2

### 2.8 Summary

The methods and techniques presented here are basic tools for understanding and modeling the spatial variation in semiconductor processes and devices. Such models are important for improving both process development and circuit design. By factoring or deconvolving a raw dataset into wafer-level, die-level, and wafer-die interaction terms, the effect of different process flows and equipment drift and consumables replacement on
both the die-level and wafer-level variation can be measured. For technology CAD, these techniques represent a prospective means for detailed physical simulation model development and calibration. These results are also applicable to statistical circuit simulation.

**Figure 2.16:** Histogram Comparison and Quantile-Quantile Plot for HP Mask Set (Wafer B5)

From plots similar to Figure 2.16 the mean and standard deviation of the wafer-level and die-level variation components, or more appropriately the actual distribution, can be used as input to Monte-Carlo or macro-model based circuit simulators in order to predict circuit performance variation caused by layout factors and to predict the expected performance
spread across the wafer as a result of wafer-level variation. In this way, designs can be simulated and improved to create variation insensitive designs.

Several extensions to previous work on spatial process/device variation have been presented. Spline and filter methods for wafer-level modeling have been proposed, and comparisons with an analytic regression approach have been made. The use of the autocorrelation function to aid in wafer-level modeling is also suggested. In addition, a physical cross-validation method has been suggested to build confidence in the structure of wafer-level models. For periodic pattern-dependent die-level variation extraction via the Fourier transform, we have described error analyses that show the importance of sufficient die sampling on a wafer for accurate extraction. New spline and frequency filtering methods for identifying wafer-die interaction effects have also been proposed. Finally, correlogram and residual tests for evaluating variation modeling effectiveness have been described. These elements have been combined into an integrated methodology for effective modeling of spatial variation.
Chapter 3

Rapid Variation Assessment and Diagnosis

3.1 Introduction

The techniques presented in the previous chapter enable accurate variation decomposition and analysis. While the accuracy of the technique is quite good, the amount of data needed is very large often requiring sampling nearly every die on a wafer. This data budget is acceptable during equipment ramp-up and for research and exhaustive equipment modeling, but for use as a process monitor or as a diagnosis technique, methods which economize on the amount of data are required. This chapter presents alternative methods for variation decomposition and analysis which only need a fraction of the data bandwidth required for the full techniques mentioned in the previous chapter. After describing the techniques for rapid decomposition into wafer-level, die-level, wafer-die interaction, and random residual terms, an example is presented.

3.2 Wafer-Level Variation

An approximate estimator for wafer-level variation is the single structure method (SSM). In this method, the same structure, usually near the center of the die, is measured for all die on the wafer as shown in Figure 3.1. This typically results in approximately 75 to 100 measurements per wafer. The data is then interpolated to obtain values across the entire wafer.
Figure 3.1: The single structure method samples the same structure on each die across all dies on a wafer typically requiring 75-100 samples per wafer. Most automated metrology tools can measure a wafer using this strategy in about five minutes.

The SSM method can fail to accurately estimate the wafer-level variation [38] especially if the site selected for the SSM method is not near the center of the distribution of all sites across a particular die (see Figure 3.2). A better choice for a rapid wafer-level estimator is to measure a “high” spot and a “low” spot on each die and average the two before
interpolation. While this results in doubling the number of samples per die, the estimator is more robust due to additional averaging.

Figure 3.2: The wafer-level variation estimator using the single structure method (SSM) can yield erroneous results especially if a site is selected which does not modulate with the wafer-level variation.

3.3 Die-Level Variation

The die-level variation can be approximated by position averaging. This estimate was introduced in [8] and is computed by averaging all measurements from the same structure measured on each die (see Figure 3.3). The process continues for all structures of interest. The grand mean is removed to force the die-level variation to center around zero.

\[ P_{PAVG}(x, y) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i, y_j) - \mu \]

x,y coordinates refer to the position within the die for the jth row and ith column die.

Figure 3.3: A rapid estimator for the die-level variation can be computed using position averaging. Position averaging is computed by computing the average of all measurements for the same structure across all die and then subtracting the grand mean.
3.4 Wafer-Die Interaction Terms

In many cases, the wafer-die interaction term can be approximated by assuming the interaction terms are proportional to the die-level variation:

\[ f_{\text{CROSS}}(i, j) \propto \alpha(i, j) \cdot f_{\text{DLV}}(i, j) \]  

(3.1)

where \( f_{\text{CROSS}} \) are the wafer-die interaction terms, \( f_{\text{DLV}} \) is the die-level variation, \( i,j \) are the \( j \)th row and \( i \)th column die, and \( \alpha \) is constant within each die. The coefficient \( \alpha \) is estimated by linear regression. For each die, a linear model of the remaining systematic terms (i.e. \( f_{\text{RAW}} - f_{\text{DLV}} - f_{\text{WLV}} \)) versus the die-level components is built. In such a model, the \( \alpha \) coefficient reduces to the correlation between these two terms.

As the definition in (3.1) indicates, \( 1 + \alpha \) allows the wafer-die interaction terms to be visualized as the extent to which the die-level variation pattern is compressed or expanded at each die across the wafer. For \( 1 + \alpha > 1 \), the wafer-die interaction terms are expansive, and for \( 1 + \alpha < 1 \) the interaction terms are compressive.

3.5 An Example

Figure 3.4 shows the raw ILD thickness variation for a wafer in a CMP process with every site on every die sampled optically to obtain over five thousand measurements. On the data from this wafer, the exhaustive techniques described in the previous chapter were used. On the same wafer, all sites on five die near the center and four die near the edge were sampled (approximately 150 measurements) in addition to one site near the center of all die (approximately 75 measurements) on the wafer. On this data set, the rapid decomposition techniques were applied and compared to the more exhaustive techniques. In this experiment, a split was done on polishing pad type (IC-1400 vs. IC-2000) a key consumable variable.
Figure 3.5 shows a comparison of the SSM method versus the down-sampled moving average (DSMA) method discussed in Chapter 2 for a sample wafer from each pad. Figure 3.6 shows the die-level variation for a sample wafer from each pad split. The wafer-level estimators and die-level estimators are qualitatively very similar. Figure 3.7 shows a more quantitative comparison of the wafer-level estimators and the die-level estimators. For the die-level estimators, the two estimators are very well correlated except that the slope is not one. This error may result from only sampling over five interior die to obtain the die-level estimator. For the wafer-level estimator, however, the results are not as good. Surely, the discrepancy is related to the limitations of the single structure method (see Section 3.2).
Figure 3.5: The wafer-level variation estimated using the DSMA technique (a,c) and the SSM technique (b,d). The extracted wafer-level variation is shown for one wafer from each pad split.
Figure 3.6: The die-level variation estimated using the FFT techniques (b,d) and using the position averaging techniques (a,c). Data is shown for a representative wafer from each pad split.

Figure 3.7: (a) A comparison of the single-structure method versus the down-sample moving average method and (b) the position averaging method versus the FFT method.
Figure 3.8 shows both the interaction terms for a sample wafer from each pad-type extracted using the FFT techniques described in Chapter 2, and the $1 + \alpha$ surface as a function of position. We see that both the extracted wafer-die interaction terms (Figure 3.8a,b) and the $1 + \alpha$ coefficient surfaces (Figure 3.8c,d) share some similarities. Since the multiplicative model in (3.1) is assumed to be proportional to the wafer-die interaction terms, the exact value of the terms can only be known to within a constant; however, the basic dependence and features of the wafer-die interaction terms can always be observed. An important observation results from examination of the $1 + \alpha$ surface: by comparing the surface to the wafer-level pattern (Figure 3.5), we see that the wafer level trends may indeed contribute to the expansion or compression of the die-level pattern that is captured in the wafer-die interaction terms.
The remaining terms are considered to form the residual components. Using the more sophisticated estimators, the residuals were approximately +/- 18nm compared to +/- 90nm for the simplified techniques. Surely, there are some remaining systematic terms in the residual components, and this is fundamentally one of the limitations of using the simplified techniques.
Chapter 4

Variation Modeling: Investigative Modeling

4.1 Introduction
The focus of this chapter is developing modeling methods to capture the functional relationships between variation and layout factors as well as other exogenous variables such as time, equipment, consumables, and process factors. It is important to note that we are especially interested in modeling systematic sources of variation rather than collecting distributional information about variational sources. While useful, distributional information only bounds the variation, it does not enable determining cause and effect relationships between variation and layout factors. These relationships are enabled through the use of short-flow test vehicles, TCAD tools, and statistical analysis.

In Chapters 2 and 3, elements of variation decomposition were presented. As shown in Figure 4.1, variation modeling coexists with variation decomposition. Variation modeling seeks to model the components defined through variation decomposition as a function of underlying factors and root causes. Primary consideration, however, is always given towards weighting of modeling efforts. For example, variation assessment might reveal that the variance in a parameter of interest is mainly a within-wafer effect while within-die effects are small. In this case, the variation modeling phase would spend a proportionally large time developing cause and effect relationships between wafer-level variation and underlying root causes. This philosophy, however, is tempered by constraints from the variation impact stage in which the majority of the modeling effort is spent analyzing the component which has the greatest impact (which may not necessarily be the largest component).
Variation modeling usually consists of a two tiered approach: investigative modeling and analytical modeling. In investigative modeling, there is relatively little knowledge and experience regarding the functional forms and factors which the variation model should capture. There is little indication which responses should be regressed to which factors and over what functional form. For example, it may not be known whether the relationship between metal-CD and line spacing is linear, quadratic, or higher order or whether line spacing alone is sufficient to explain a large majority of the variance in the underlying variable. This situation is especially common during ramp up of a new process or for electrical variables which only couple to physical variables via complicated or largely unknown equations (e.g. $\beta$ for BJTs).

**Figure 4.1:** Variation modeling relies upon the results of the variation assessment phase. There are two types of variation modeling approaches: investigative modeling and analytical modeling. Often the results of an investigative model will form the basis of or lead to the development of an analytical model.

Analytical modeling is often used in cases where there is good reason to know that the parameter variation should follow some functional form and should involve only a select number of variables. Analytical modeling is also used when individual tests masks have been constructed for probing the individual main effects of only one or two factors at a time. Analytical modeling is also especially useful in cases where a semi-empirical model
or physically based model is available, but needs to be calibrated. In many cases, the results of investigative modeling will form the basis of or lead to the development of an analytical model.

This chapter discusses investigative variation modeling approaches while the next chapter illustrates analytical variation modeling. In this chapter, polysilicon critical dimension (CD) variation is used as an application of investigative variation modeling in order to identify the key layout factors involved in poly-CD intra-die variation and to develop first-order semi-empirical models for poly-CD variation. The next chapter uses pattern dependent variation in oxide CMP processes as a case study in analytical variation modeling.

4.2 Motivation

Poly-CD variation is especially important. Variation in the polysilicon critical dimension, or poly CD, translates directly into MOS transistor channel length variation. Current lithography and etch technology can typically achieve wafer-scale line width uniformity of approximately 5% (measurements of the same structure within each chip across the wafer). Design typically assumes that the variation within any one chip will be smaller than this. However, measurements of supposedly identical structures within the same die reveal variations on the order of 15% [6]. Clearly, additional physical effects are coming into play at this scale, such as pattern dependencies during etch or systematic lens distortions in photolithography. The modeling of spatial variation not only across the wafer, but also intra-die, is of critical concern for circuit performance and manufacturability.

4.2.1 Experimental Methodology

Since a large volume of statistically significant data is required, direct optical measurement of poly-CD is impractical. An early attempt at using optical/SEM techniques resulted in an excessively low throughput [6]. For this reason, electrical measurement
techniques are desired. An independent experiment based on simulation and physical measurements revealed that the majority of $I_{\text{sat}}$ variation is attributable to poly-CD variation as opposed to threshold voltage variation and substrate doping concentration distributions. Thus, $I_{\text{sat}}$, the measured drain current when the gate and drain are connected to the power supply and the source is grounded (for NMOS), is used as a metric of poly-CD variation.

A set of 27 NMOS transistors with a nominal channel length of 0.35\,\mu m and various layout factors including gate width, geometric orientation, spatial location, and presence or absence of neighboring structure were selected from a standard test vehicle (see Table 4.1). Every die on twelve wafers from two different lots were probed to yield over 16,000 observations.

<table>
<thead>
<tr>
<th>Number of Transistors</th>
<th>Geometric Orientation</th>
<th>Spacing</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Horizontal</td>
<td>Fingered</td>
<td>14 \mu m</td>
</tr>
<tr>
<td>7</td>
<td>Horizontal</td>
<td>Isolated</td>
<td>14 \mu m</td>
</tr>
<tr>
<td>1</td>
<td>Horizontal</td>
<td>Isolated</td>
<td>9 \mu m</td>
</tr>
<tr>
<td>2</td>
<td>Horizontal</td>
<td>Stacked</td>
<td>14 \mu m</td>
</tr>
<tr>
<td>4</td>
<td>Vertical</td>
<td>Fingered</td>
<td>14 \mu m</td>
</tr>
<tr>
<td>7</td>
<td>Vertical</td>
<td>Isolated</td>
<td>14 \mu m</td>
</tr>
<tr>
<td>1</td>
<td>Vertical</td>
<td>Isolated</td>
<td>9 \mu m</td>
</tr>
</tbody>
</table>

### 4.2.2 Wafer level model

A regression approach is used to extract the wafer-level model. In this estimator, the wafer-level variation is assumed to be a smooth function of $x$, $y$, $r$, $\theta$, or other relevant coordinates. The estimator used here is of the form:

$$f_{\text{WLV}} = \alpha_0 r + \alpha_1 r^2$$

(4.1)

where $r$ is the radial distance to the center of the wafer and standard least squares regression to the available data is employed.
4.2.3 Die level and Wafer-Die Interaction Term Model
A “VarNOVA” procedure (VARiation modeling using aNalysis Of VAriance) has been developed to model die-level variation and wafer-die-interaction components. The process begins by developing an ANOVA model for each die measured as a function of the designed layout factors (e.g., Table 4.1). The model coefficients as well as the spatial coordinates of each die are assembled together. These model coefficients are then regressed onto a polynomial function of the spatial coordinates of each die usually using a MANOVA technique. The constant terms (which are independent of the die spatial coordinates) thus form the die-level (or intra-die) component of the model, while those which are a function of the die coordinates form the wafer-die interaction components. The remaining terms become the residual components.

![Image of model coefficients and interaction terms]

Figure 4.2: The VarNOVA procedure

4.2.4 Poly-CD Data Analysis
Figure 4.3 shows the (normalized) raw data for one particular wafer. The smooth, low-frequency characteristics of the wafer-level components can be seen as well as local bumpiness associated with die-level and wafer-die interaction components of variation.
Figure 4.4 shows the extracted wafer-level models for four measured wafers. The upper two wafers come from Lot “A” while the lower two wafers come from lot “B”. In all the wafers, there is a clear radial dependence as well as a linear trend visible. Also, significant lot-to-lot and wafer-to-wafer variation is visible.

For each measured die, an ANOVA model:

\[
Id_{\text{SAT}} = \mu + \left\{ \begin{array}{c}
\alpha_F \\
\alpha_I \\
\alpha_S \\
\beta_H \\
\beta_V \\
\gamma_{\text{WIDE}} \\
\gamma_{\text{NARROW}}
\end{array} \right\}
\]

(4.2)

is developed. Since the design was unbalanced and research revealed that interaction terms (e.g. $\alpha_F \times \beta_H$) added little to the quality of the model, interaction terms were neglected. In (4.2), $\mu$, $\alpha_F$, $\alpha_I$, $\alpha_S$, $\beta_H$, $\beta_V$, $\gamma_{\text{WIDE}}$, and $\gamma_{\text{NARROW}}$ are constants which are fitted to the data.
In this type of model, the expected $I_{dsat}$ value for a particular transistor with a given set of layout factors is determined by selecting the appropriate constant term from each brace and adding each term together. For example, if a transistor is horizontally oriented, isolated, and wide, then the expected $I_{dsat}$ value is $\mu + \alpha_i + \beta_H + \gamma_{WIDE}$. Also note that all of the models are fitted to the data with the wafer-level model removed.

A typical ANOVA model (normalized) for a die near the center of the wafer is:

$$I_{d_{SAT}} = 0.409 + \left\{ \begin{array}{c} 9.375 \\ 10.630 \\ 1.254 \end{array} \right\} + \left\{ \begin{array}{c} -3.162 \\ 3.162 \\ 1.641 \end{array} \right\} + \left\{ \begin{array}{c} -1.641 \\ 1.641 \end{array} \right\}$$  \hspace{1cm} (4.3)

and the ANOVA table for this particular die is shown in Table 4.2. The ANOVA table reveals that spacing (fingered vs. isolated vs. stacked) and geometric orientation (horizontal vs. vertical) are the most significant layout factors while channel width is not signifi-
cant as judged by the Pr(F) column (under a few assumptions of normality, there values indicate the probability that the observed differences between groups could have arisen by chance alone). For this reason, the channel width layout factor is not included in any of the models presented below.

![Figure 4.5: The Spatial Dependence of the Model Coefficients for Wafer 21, Lot “A”](image)

**Table 4.2: ANOVA Table for Wafer 21, Lot “A”, Die 44.**

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>Sum of Sq.</th>
<th>Mean Sq.</th>
<th>F-value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>spacing</td>
<td>2</td>
<td>6964</td>
<td>3482</td>
<td>81.41</td>
<td>0.0000</td>
</tr>
<tr>
<td>geom. orient.</td>
<td>1</td>
<td>745.8</td>
<td>745.8</td>
<td>17.44</td>
<td>0.0003</td>
</tr>
<tr>
<td>width</td>
<td>1</td>
<td>56.8</td>
<td>56.8</td>
<td>1.32</td>
<td>0.2614</td>
</tr>
<tr>
<td>Residuals</td>
<td>22</td>
<td>940.961</td>
<td>42.77</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 4.5 shows plots of the model coefficients, μ, α_F, α_I, α_S, β_H, and β_V, as a function of die spatial location on the wafer. The MANOVA procedure described in Section 4.2.3 is used to decompose the wafer-level residuals (the difference between the raw observed variation and the extracted wafer-level variation) into die-level and wafer-die interaction components models and thereby capture the majority of the systematic variation visible in Figure 4.3. The results of the decomposition for one particular wafer are
detailed in (4.4) where the coefficients have been regressed onto a polynomial function of spatial position (radius from the center of the wafer, \(x\) position from the center of the wafer, \(y\) position from the center of the wafer, and the product of \(x\) position and \(y\) position from the center of the wafer.

\[
\begin{bmatrix}
\mu \\
\alpha_F \\
\alpha_I \\
\alpha_S \\
\beta_H \\
\beta_V
\end{bmatrix} =
\begin{bmatrix}
0.581 & -0.708 & -0.622 & 0.002 \\
-0.067 & -0.548 & -0.369 & 0.085 \\
0.355 & 0.289 & 0.275 & 0.181 \\
-0.287 & 0.259 & 0.094 & -0.266 \\
0.476 & 0.102 & -0.106 & 0.035 \\
-0.476 & -0.102 & 0.106 & -0.035
\end{bmatrix}
\begin{bmatrix}
r \\
x \\
y \\
x \cdot y
\end{bmatrix} +
\begin{bmatrix}
1.863 \\
8.470 \\
-10.410 \\
1.941 \\
-3.019 \\
3.019
\end{bmatrix}.
\] (4.4)

As mentioned in Section 4.2.3, the terms which depend on spatial position become the wafer-die interaction components while the constant terms, i.e., those independent of spatial position, become the die-level component of the model. In this way, the die-level component of the model is common to all measured dice on the wafer. Also, the correlation coefficient for each component of the model shown in (4.4) is listed in Table 4.3.

**Table 4.3: The Correlation Coefficient for the Generated VarNOVA Model**

<table>
<thead>
<tr>
<th>(\mu)</th>
<th>(\alpha_F)</th>
<th>(\alpha_I)</th>
<th>(\alpha_S)</th>
<th>(\beta_H)</th>
<th>(\beta_V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.66</td>
<td>0.73</td>
<td>0.71</td>
<td>0.52</td>
<td>0.51</td>
<td>-0.51</td>
</tr>
</tbody>
</table>

Developing a model similar to (4.4) serves two enabling purposes which are not easily accommodated through simpler analysis techniques. First, by developing models for all measured wafers and comparing coefficients and spatial dependencies, hypotheses regarding lot-to-lot and wafer-to-wafer variation can be tested. Furthermore, by looking at the spatial dependence of each factor type, geometric orientation versus spacing in this particular case, possible different physical and spatial dependencies can be highlighted. For example, (4.4) indicates that isolated features have a positive dependence on radial position while the stacked and fingered components have a negative dependence on radial
position. The relatively low correlation coefficients illustrated in Table III may result from either a large random component inherent in the data and/or a more complex spatial dependence which is only properly formulated in a non-parametric sense. Nevertheless, the MANOVA model seems to capture important first-order dependencies.

4.2.5 Summary
For the poly-CD data variation observed in this work, spacing, or distance to nearest structure, as well as geometric orientation are the dominant layout factors while channel width is not a significant component. However, further research is needed to understand the role of other layout factors such as poly density. Also, the role of channel width cannot be ruled out completely since only two levels are explored here and the design space for these two levels are severely unbalanced (see Table 4.1).

Finally, the VarNOVA methodology allows simple first-order models to be developed for poly-CD variation as manifested in $I_{sat}$ variation. Furthermore, the models developed using this methodology allow the exploration of layout/pattern effects as a function of spatial position. The fingered structures seem to illustrate a different radial dependence compared to the isolated structures; this is an intriguing result and further research can now be targeted to explore this effect which might have otherwise been buried in noise.
Chapter 5

Analytical Modeling

5.1 Introduction
Since the diversity of problems encountered in analytical modeling is very large, completely generalizing analytical modeling approaches is extremely difficult. The fundamental physics and resulting approaches to modeling pattern dependent photolithography variation is quite different from chemical mechanical polishing. However, analytical modeling generally involves special test masks, metrology techniques, and analysis tactics. In recognition of this, a case study of analytical modeling of pattern dependent ILD thickness variation in oxide CMP processes is presented in this chapter. This case study describes and uses the CMP Characterization mask set with accompanying metrology and analysis techniques.

After discussing previous work in CMP pattern dependent characterization, the CMP characterization masks are presented in Section 5.3, and the metrology tools and techniques for which the masks were designed are described in Section 5.4. An experimental application of the test masks is presented in Section 5.5, in which the polishing performance of two different pads are evaluated. The analysis methods and procedures are presented and illustrated for this pad experiment in Section 5.6.

5.2 Previous Art
In this paper, we present four masks for characterizing and modeling pattern dependent variation in CMP processes, consumables, and tools. Using these masks, we present methods for the rapid characterization, empirical modeling, and comparison of pattern dependencies as a function of processes, consumables, or equipment options. This is
achieved in two ways. First, each mask is targeted toward an individual source of pattern dependent variation. To this end, four separate single-layer masks have been designed to probe structure area, pattern density, line pitch, and structure aspect ratio effects, respectively. Second, the masks support simplified metrology tools and techniques including optical film thickness and profilometry measurements.

Several masks have been developed for analyzing pattern dependent variation. Warnock [47] and Hayashide, et al. [48] used a simple mask composed of a set of lines and spaces varied across the die. Renteln, et al. [49] also used a set of gratings but over very large feature scales (1-10mm). SRAM or gate array cells are sometimes used [50, 51]. Burke [41] used a set of two test masks but does not describe their design or construction. With a few exceptions, these masks and accompanying analysis techniques are not discussed fully or at all, and appear primarily targeted at testing specific models rather than to aid in characterizing or generating models.

Chang et al. [5] and Stine et al. [52] have demonstrated the use and analysis of test masks for screening experiments designed to investigate and identify the potential factors contributing to ILD thickness non-uniformity. These masks, however, were designed for electrical probing; as a result, their use requires several masking steps and is often restricted by the availability and throughput of probing equipment. Substantial effort is also required to analyze electrical probe data, including the conversion of electrical data to ILD thickness information. Electrical test masks do provide valuable complementary information to the simpler test masks described in this paper: small linewidth and spacing features can be used and combinations of layout factors that more closely mimic realistic circuits can be constructed.
5.3 Mask Descriptions

The CMP Characterization Mask Set (Figure 5.1) is designed for rapid CMP consumable, process, and tool characterization and evaluation. Each mask is designed to produce a 1.2cm x 1.2cm die, but a smaller die can be generated by scaling the layout. The first mask, the **area mask**, (Figure 5.1a) has patterned structures with areas ranging from 10x10 µm² to 3x3 mm² across a variety of pattern densities achieved by altering the fill pattern inside each structure. In addition to the area structures, there are also structures to test the role of geometric orientation (horizontal lines versus vertical lines).

The **pitch mask** is the second mask (Figure 5.1b). The density of each structure is fixed at 50% (equal linewidth and linespace), and the pitch is varied from 2 µm to 1000 µm for a total of 36 structures for each die. With the exception of the structures with a pitch less than 20 µm, each structure is 2mm by 2mm in size. Features with a pitch of 20 µm or less are lumped into a single 2mm x 2mm structure. There are also spatial replicates for many of the structures so that pitch effects can be separated from spatial location effects.

In the **density mask**, the third mask (Figure 5.1c), the pattern density (the ratio of raised metal area in each structure to the total area of each structure) is varied systematically from 4% to 100% from lowest in the lower left corner to greatest in the upper right corner while the pitch of each structure was fixed at 250 µm. A total of 25 structures, each 2mm x 2mm, are arranged in a 5x5 grid, in addition to a border region which extends 1mm from the edge of the 25 structures to act as a buffer. Without the border, structures with the lowest density would contact structures with the highest density (on a neighboring die) and interact strongly.
The fourth mask, or the aspect ratio (perimeter/area) mask (Figure 5.1d), is designed to explore the role of aspect ratio (the ratio of the length of the structure to the width of the structure) and the ratio of perimeter to area. This mask targets any systematic edge/corner effects which may be present. A total of eight structures replicated twice are designed. The area of each structure is fixed at approximately 1mm$^2$ and the ratio of width to length is fixed at 1:1, 1:4, 1:14, 1:16, 1:50, 1:62, 1:82, and 1:100. Each set of 16 structures is replicated six times across the die but with different spacings between structures ranging linearly from 10 μm to 60 μm for a total of 96 structures.

Figure 5.1: The CMP Characterization Masks: the area, pitch, density, and aspect ratio masks are shown in Figures 5.1a-d respectively. Each die is 12mm x 12mm
5.4 Metrology Techniques

In this section, the application of several metrology tools and techniques are discussed. Specifically, particular attention is directed towards optical interferometry and profilometry techniques as well as Atomic Force Microscopy (AFM) and SEM (Scanning Electron Microscopy).

Optical interferometry is a commonly used technique [57] for directly measuring film thicknesses. It offers reasonably high throughput as well as an absolute thickness measurement assuming the tool is properly calibrated. For the characterization masks, either five die, nine die, or a full wafer sampling schemes are useful. In a five die strategy, five die near the center of the wafer (which is typically more uniform than elsewhere on the wafer) are measured. This is useful for quick analysis and initial model development, but does not provide any information about edge or spatial effects across the wafer. In a nine die sampling scheme, one die at each edge of the wafer (top, left, right, and bottom) and at the same radius are sampled in addition to five die from the center of the wafer. This technique can provide some information about edge effects. Finally, a full-wafer scenario can be used in which every die is sampled on the wafer. While the throughput of this technique is quite low, more powerful analytic techniques such as those described in [53] can then be used. Automated optical metrology is limited to structures with linewidths greater than 10 μm. For smaller structures (down to about 4 μm), optical metrology can still be used but only in manual model and with less reliable results.

The center of each structure above the metal (or above the field in the case of a trench process) is measured for the specified die, resulting in 20-30 measurements per die. The aspect ratio mask is an exception; in this case each structure from only one of the six spac-
ings regions is sampled (16 measurements), and then one structure from each spacing region is also sampled across all spacings (five additional measurements) for a total of 21 structures. If desired, the thickness in the field next to the metal lines can also be measured to provide planarization information, and knowledge about the polishing time can be used to compute removal and planarization rates.

In profilometry, a sharp stylus is dragged across the surface of interest and deviations in the stylus are measured [57]. Profilometry can be used in 3-D mode to generate an entire die map with high throughput or in 2-D mode to generate planarization information over centimeter scale distances. In profilometry, measurements are susceptible to stage tilt and bias as well as wafer bow and warp. This problem can be compensated by using a combined optical/profilometry technique: several points are selected on the die which are measured using both optical interferometry and profilometry measurements. The surface formed by the measurements from profilometry is forced to match the surface formed by the optically measured values; the difference between these two surfaces is called the correction surface. The effectiveness of this technique is dictated by the number of measurements which make up the correction surface: using only five measurement sites to form the correction surface enables correction for linear deviations.

AFM and SEM [57] are also useful metrology tools for use with the characterization masks. In each case, detailed information about planarization can be obtained, and small structures which cannot be measured using optical interferometry can also be examined. Since the effective throughput of these techniques is extremely low, the application of these tools is limited to only a few key measurements. An additional drawback of SEM techniques is that they are destructive.
Finally, a simple but effective approach should not be neglected: visual inspection and interpretation of fabricated wafers. In certain circumstances, differences between structures and wafers can be discerned visually by looking for subtle color variations between each case, and a color chart [57] can be used to roughly quantify thickness differences. Although simple, this technique can be quite useful in identifying gross differences associated with particular consumable, process, and tool choices during rapid evaluation or optimization.

5.5 Experimental Results

In this section, a pad comparison experiment and raw data are described to illustrate the use and application of the characterization mask set. The analysis and modeling of this data is discussed in the following section. For this experiment, a back-end process is used in which field oxide is initially deposited followed by metal deposition, pattern, and etch. After depositing the ILD (inter layer dielectric), the wafers are polished to achieve uniform surface heights. A detailed process flow used in this study can be found in Figure 5.2.

<table>
<thead>
<tr>
<th>Table 5.1: The Polishing Parameters Used in the Pad Comparison Study.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tool Settings</strong></td>
</tr>
<tr>
<td>Down Force</td>
</tr>
<tr>
<td>Table Speed</td>
</tr>
<tr>
<td>Quill Speed</td>
</tr>
<tr>
<td>Back Pressure</td>
</tr>
<tr>
<td>Slurry Flow Rate</td>
</tr>
<tr>
<td><strong>Consumable Settings</strong></td>
</tr>
<tr>
<td>Slurry</td>
</tr>
<tr>
<td>Carrier</td>
</tr>
<tr>
<td>Carrier Insert</td>
</tr>
<tr>
<td>Conditioning Technique</td>
</tr>
</tbody>
</table>
Field Oxide Deposition
0.8μm PETEOS

Metal Stack Formation
25nm TiW
0.75μm Al/0.5% Cu
25nm of TiN

Metal Pattern and Etch

ILD Deposition

CMP
with IC-1400 Pad

CMP
with IC-2000 Pad

Metrology

Optical Measurements
1 Wafer/Pad Split
Randomly Selected Wafer
5 Die Near Center Mea-

Profilometry Measurements
1 Wafer/Pad Split
Randomly Selected Wafer

Figure 5.2: The short-flow process flow used in the pad comparison study.
For the polishing process, the wafers are split on polishing pad type. Half of the wafers are polished with an IC-1400 polishing pad while the other half are polished with an IC-2000 polishing pad, both from Rodel, on a IPEC/Planar 472 polishing tool. The polishing parameters for the pad experiment can be found in Table 5.1. Several dummy wafers, or wafers which contain only a thick blanket oxide layer, are polished on each pad type to “break-in” the pad, i.e., to stabilize the removal rate of the pad.

Figures 5.3a-d show ILD thickness, measured on a Prometrix FT-650- an optical thin film measurement tool, versus area, pitch, density, and aspect ratio respectively. Each data point in Figure 5.3 represents the average of five die near the center of the wafer. Note that Figures 5.3a,d are shown as ΔILD thickness versus area and aspect ratio. ΔILD represents the component of ILD thickness variation that can be explained by area or aspect ratio alone. The component of ILD thickness variation which could be explained by differences in pattern density was removed prior to determination of the area or aspect ratio contribution; this procedure is described in detail in Section 5.6.
Figure 5.3: ILD thickness versus structure area, pitch, pattern density, and aspect ratio for each pad type. Note that for the area and aspect ratio data, Δ ILD thickness values are listed since the effect of pattern density has been removed from the data.
Figures 5.4a-h show the measured profilometry traces for each mask measured using a Tencor P-22 profilometry for a die near the center. The leveling techniques discussed in
the previous section were used to remove wafer bow and warp and stage bias. A linear correction surface was used as discussed in the previous section. In all cases, the origin of the profilometry measurements was chosen as close as possible to the lower left corner of each pattern. The differences between the two pads, especially for the area mask, are discussed and explained in the next section.

In addition to measuring the post-CMP ILD thickness, we also selectively measured the pre-CMP (i.e. immediately after deposition) ILD thickness. From these measurements and based on the fact that we measured five die near the center of the wafer where incoming deposition variation is smallest, we concluded that deposition variation is small (typically +/- 10nm); thus, post-CMP measured thicknesses did not have to be corrected for incoming film deposition variation.

**5.6 Pattern Dependent Variation Model Generation**

In this section, the use of the characterization masks for pattern dependent variation modeling in CMP processes is presented. Several modeling methodologies, spanning a range of applicability and sophistication, are presented for each mask and applied to data obtained from the pad experiment described in Section 5.5. The resulting models are especially useful for verifying physical models, for process optimization, or for studying the impact of variation on circuit performance or manufacturability.

In order to generate semi-empirical models, two wafers were randomly selected for each mask—one from each pad split. Five die from the center of each wafer were averaged together to partially suppress within-wafer effects. Models of ILD thickness variation were formed as a function of area, pitch, density, and perimeter/area (aspect ratio).
5.6.1 Pattern Density Modeling

As reported in the literature [41, 42, 5, 52, 58] and apparent from a visual inspection of the data (Figures 5.3, 5.4), the ILD thickness is quite sensitive to pattern density. A major obstacle to modeling pattern density dependencies in CMP rests with finding a suitable and compact definition for a density metric which not only yields a good fit to the available data, but is also physically intuitive. [59, 60]

**Figure 5.5:** A simplified example to aid in the definition of pattern-density.

An example helps to illustrate the definition of pattern density. Figure 5.5 shows a simple cross section through a fictitious test structure composed of two 1 mm wide metal lines separated by 1 mm and a 5 mm line which is separated from the 1 mm lines by 3.5 mm. Since the lines are very wide, we can assume that the deposition profile can be approximated by the metal profile. In this example, a 1.5 μm layer of oxide was deposited. In many situations the deposition is conformal and not as shown in Figure 5.5, and the oxide profile cannot always be approximated by the metal profile; this is most evident in tight pitches or small spaces. For this reason, computations of pattern density also depend upon accurate deposition profiles or models, and deposition parameters, tools, and materials are an important integration/modeling issue in CMP [60]. Fortunately, the linewidth and space...
(except for the pitch mask) are all greater than 10 microns, and approximating the deposition profile with the metal profile is a reasonable approximation.

According to the model proposed by Stine, et al. [59], the relationship between ILD thickness and pattern density can be expressed as:

\[
  z = z_0 - z_1 - Kt + \rho z_1 \quad \text{for } t > \frac{z_1}{K}
\]

where \( z \) is the ILD thickness referenced from the top of metal regions, \( z_0 \) is the amount of dielectric deposited before CMP, \( z_1 \) is the as-deposited step height, \( K \) is the removal rate of blanket or unpatterned wafers, \( t \) is time, and \( \rho \) is pattern density.

**Figure 5.6:** The definition of planarization length. Typical planarization lengths are on a scale of several mm

In the literature, much discussion has been generated over what is the planarization length for any particular CMP process. Figure 5.6 illustrates a definition of planarization distance. If one seeks to planarize “vertical” oxide profiles (as shown in Figure 5.5) over
two regions with a step change in pattern density (as shown in Figure 5.6), the low density region will polish faster than the high density region with a transition ramp in between the two regions. The final oxide profile will be similar to the ramp shape shown in Figure 5.6, and the planarization length is defined to be the width of this transition ramp.

In this paper, we define pattern density at a particular location \((x,y)\) on a die as the area of all polygons inside a square region called the *density window* (see Figure 5.7) divided by the area of the density window. We call the length of a side of the density window the *interaction distance* \((id)\). An intuitive physical interpretation of the *interaction distance* is the macroscopic distance over which the pad bends and conforms to the wafer surface and is typically several mm. In the pad experiment described in Section 5.5, the IC-2000 is a stiffer pad compared to the IC-1400 and also has a longer interaction distance as revealed.

**Figure 5.7:** The definition of interaction distance shown on an example layout.
below. It can also be shown for pattern density computed using a square density window that the interaction distance is identical to the planarization length.

Regardless of the techniques used to compute pattern density, a procedure is needed for optimizing the choice of the interaction distance parameter. Three ways are available for determining the interaction distance: (1) via direct measurement of the planarization length, (2) via a slope regression method, or (3) via a maximum $R^2$ method. Direct measurement of the planarization length is by far the quickest and most direct method of determining the interaction distance. Unfortunately, none of the structures on any of the characterization masks are large enough or contain a large step change in density; thus, this technique was not used in this experiment.

In the slope regression method, the slope of a line fit of ILD thickness to pattern density for a particular mask is computed over a wide range of interaction distances. Thus, if the pattern density is computed at the proper interaction distance, the slope of a regressed line of ILD thickness to pattern density should equal the as-deposited step height. If the regressed slope does not equal the measured as-deposited step height, the procedure continues on a different choice of interaction distance until convergence is obtained.

Figure 5.8 (a,c) shows a plot of the regressed slope versus interaction distance for data from the density mask and the area mask. The amount of dielectric deposited ($z_0$) was known with considerable certainty, but the as-deposited step height, $z_I$, could only be estimated. Fortunately, data from a similar process was available and the amount of pattern sensitivity in $z_I$ is small. Based on this information, $z_I$ was estimated at 0.83μm +/- 0.01μm. As Figure 5.8 shows, the estimated interaction distance is 3.2-3.4mm for the IC-1400 pad and 4.2-4.4mm for the IC-2000 pad as measured from the density mask data.
The area mask data indicates that the interaction distance is 3.6mm for the IC-1400 pad and 4.2mm for the IC-2000 pad. Note that the area mask is more accurate in detecting the interaction distance compared to the density mask. Since the density mask was designed with gradual gradations in pattern density from one region to the next, this result is not surprising. The area mask has density values spread more randomly across the mask. This unplanned design feature results in greater discrimination of the interaction distance.

In the maximum $R^2$ method [60, 61], the interaction distance is determined by regressing a linear function of pattern density to ILD thickness for several different values of the interaction distance. The $R^2$ value, a measure of the quality of the model, is computed for each value of interaction distance. The interaction distance which maximizes $R^2$ is selected. We have found that the maximum $R^2$ method is unreliable and imprecise and can yield misleading results. Figure 5.8 (b,d) shows a plot of $R^2$ versus interaction distance for data from the density mask and the area mask. We note that the slope regression method is considerably more precise in detecting the interaction distance, and for the area mask, the interaction distance extracted using the $R^2$ method is significantly different than the slope regression method. This is due to the presence of an underlying design factor, in this case area, confounding with density. In the density mask, there are no significant confounding design factors available to degrade the determination of the interaction distance; thus, the $R^2$ method can still be used albeit with diminished precision compared to the slope regression method.
Figure 5.8: Estimated slope of ILD thickness versus pattern density computed at different interaction distances from 3 to 4.5mm for the density mask (a) and for the area mask (c). For comparison the maximum $R^2$ method is shown in (b) and (d) for the density mask.
<table>
<thead>
<tr>
<th>Model</th>
<th>ILD</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density Mask</strong></td>
<td>ILD=0.83p+0.51 [μm] (R² = 0.993) (id = 3.2-3.4mm)</td>
<td></td>
</tr>
<tr>
<td>IC-1400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(intercept)</td>
<td>0.51</td>
<td>0.0078</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.0145</td>
</tr>
<tr>
<td><strong>IC-2000</strong></td>
<td>ILD=0.82p+0.56 [μm] (R² = 0.995) (id = 4.2-4.4mm)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(intercept)</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.82</td>
<td>0.0128</td>
</tr>
<tr>
<td><strong>Pitch Mask</strong></td>
<td>ILD=2.6932x10⁻⁶pitch+0.86 [μm] (R² = 0.0088)</td>
<td></td>
</tr>
<tr>
<td>IC-1400</td>
<td>(intercept)</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>pitch [μm]</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>IC-2000</strong></td>
<td>ILD=-5.3658x10⁻⁶pitch+0.91 [μm] (R² = 0.0097)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(intercept)</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>pitch [μm]</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Area Mask</strong></td>
<td>ILD=0.80p+0.54 + ΔILD [μm] (R² = 0.89) (id = 3.6mm)</td>
<td></td>
</tr>
<tr>
<td>IC-1400</td>
<td>(intercept)</td>
<td>-0.0349</td>
</tr>
<tr>
<td></td>
<td>Area [mm²]</td>
<td>0.0367</td>
</tr>
<tr>
<td></td>
<td>Area² [mm⁴]</td>
<td>-0.0038</td>
</tr>
<tr>
<td><strong>IC-2000</strong></td>
<td>ILD=0.80p+0.60 + ΔILD [μm] (R² = 0.82) (id = 4.2mm)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(intercept)</td>
<td>-0.0365</td>
</tr>
<tr>
<td></td>
<td>Area [mm²]</td>
<td>0.0335</td>
</tr>
<tr>
<td></td>
<td>Area² [mm⁴]</td>
<td>-0.0032</td>
</tr>
<tr>
<td><strong>Aspect Ratio Mask</strong></td>
<td>ILD=0.80p+0.63 + ΔILD [μm] (id = 3.6mm)</td>
<td></td>
</tr>
<tr>
<td>IC-1400</td>
<td>(intercept)</td>
<td>-0.0019</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>-4.6x10⁻⁵</td>
</tr>
<tr>
<td><strong>IC-2000</strong></td>
<td>ILD=0.80p+0.63 + ΔILD [μm] (id = 4.2mm)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(intercept)</td>
<td>0.0058</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>-2.4x10⁻⁴</td>
</tr>
</tbody>
</table>
Table 5.2 and Figure 5.3 show the fitted model of ILD thickness versus density. The slopes of the model for each pad are essentially identical (equal to the local step height) by construction. The intercept for the IC-1400 pad data is less than the intercept for the IC-2000 pad data since the blanket removal rate of the IC-1400 pad is greater than the blanket removal rate for the IC-2000 pad and all wafers were polished for approximately the same amount of time. The interaction distance, however, is significantly different for each pad type. The larger interaction distance for the IC-2000 pad compared to the IC-1400 pad correlates well with the physical stiffness/hardness of the IC-2000 pad versus the IC-1400 pad.

The difference in interaction distance between the two pad types explains the differences in the data for the area mask shown in Figure 5.4. For an interaction distance near 3mm (IC-1400), the pattern density near the center of the larger features in Figure 5.1a is nearly 100% while for an interaction distance near 4mm (IC-2000) and for the same structure, a substantial amount of low density regions surrounding the large features are averaged together. Thus, transitions between high and low densities with the IC-2000 pad are substantially smoother and a smaller range in pattern density and ILD thickness variation can be expected compared to the IC-1400 pad.

Since the area mask set also has structures which span different ranges of pattern density, the area mask set data can also be used to check the predictive power of models generated from the masks. Figure 5.9 shows the predicted ILD thickness variation and observed ILD thickness variation measured using the combined optical/profilometry technique on the area mask. The ILD thickness variation model in Equation (1) was used with an interaction distance of 3.4mm (extracted from the density mask). As the results show,
the agreement is good. Discrepancies between the prediction and the measured values is due to lack of precise knowledge about the shape of the density window, uncompensated metrology errors, and second order effects.

\[ \text{Pad Type = IC-1400} \]

![Profilemetry measurements for a center die from the area mask polished with an IC-1400 pad (b) versus the predicted ILD thickness (a) based on computed pattern-density. Discrepancies between the model and profilometry measurements are due to lack of knowledge about the window shape, uncorrected stage bias and wafer warping, and second order effects.}

**Figure 5.9:** Profilometry measurements for a center die from the area mask polished with an IC-1400 pad (b) versus the predicted ILD thickness (a) based on computed pattern-density. Discrepancies between the model and profilometry measurements are due to lack of knowledge about the window shape, uncorrected stage bias and wafer warping, and second order effects.

### 5.6.2 Pitch Modeling

Unlike pattern density, pitch is a locally defined parameter; thus, relatively simple procedures can be used to develop a model for the dependence of ILD thickness variation on pitch. A simple regression procedure, assuming a linear or polynomial model, to the available data can be performed. A slightly more accurate model can be generated by removing
any spatial dependence (such spatial dependence may arise, for example due to any within-wafer non-uniformity localized within the measured die). The spatial dependence of the data can be estimated by generating a surface formed by replicated structures with the same pitch, such as for the 1000 μm pitch structures which were replicated six times across the die. This surface, as expressed as a function of \( x \) and \( y \) coordinates, can then be used to estimate the contribution of spatial position to each observation. Table 5.2 and Figure 5.3 shows the model and regression line and data taken from a wafer from each pad split averaged over five interior die and with spatial dependence removed. A key observation is that the dependence of ILD thickness on pitch, at least for pitches greater than 40 μm, is very slight. Also, the slope of the model does not appear to vary across pad type as indicated by the confidence intervals for the slope coefficient in Table 5.2. As before, the intercept values are different because of unequal removal rates across similar polishing times for each pad split. To compute the confidence intervals, a simultaneous confidence interval at a 95% level of confidence was used [62].

5.6.3 Minor Effects

Since structure area and aspect ratio are confounded with pattern density, a simple regression based approach, as in the pitch mask, cannot be used. For data from these masks, the portion of variation which can be attributed to pattern density must be removed first by modeling the ILD thickness variation for all structures as a function of density. The interaction distance for this density model should also be determined via one of the methods outlined in the previous subsection. Alternatively, the density model and interaction distance obtained from the density mask can be used. Once a density model has been assembled, the contribution of area or aspect ratio to ILD thickness variation can be esti-
mated by subtracting the component which can be attributed to density from the observed data. Finally, the remaining variation can be fitted to a linear or polynomial function of area or aspect ratio. Computing a model of density and area or density and aspect ratio (perimeter/area) simultaneously is inappropriate since the area, and aspect ratio are correlated with pattern density and the ILD thickness variation model (1) does not consider these effects simultaneously; thus, the model must be constructed hierarchically (i.e. the density dependence removed prior to any model development based on area or aspect ratio).

Table 5.2 and Figure 5.3 show the resulting area model for a wafer from each pad split. The density model used to remove any underlying density dependence for the area mask is also shown in Table 5.2. In order to determine the appropriate model form and order, a leave-one-out/add-one-term stepwise regression technique based on Mallow’s Cp statistic was used [62, 63]. From the stepwise procedure, a second order polynomial model emerged as the best choice for each pad split as shown in Table 5.2. Clearly, from the 95% simultaneous confidence intervals and visual inspection of the data, there is no statistically significant difference due to area between the two pad types.

The aspect ratio model for a representative wafer from each pad split is shown in Table 5.2 and Figure 5.3. Again, any underlying density dependence had to be removed; however, because the range of pattern density spanned by the aspect ratio mask is small (on the order of 20%), the estimation of the interaction distance directly from the data for the aspect ratio mask via the slope regression technique was not used. Instead, the density dependence for the area mask was assumed to apply to the aspect ratio mask data as well since all wafers were polished with the same polishing conditions and at the same time
except for the split on pad type. As the data shows, an aspect ratio effect is essentially non-existent as measured by the 95% simultaneous confidence intervals. In fact, an examination of the t-statistic for each coefficient and the overall model R² value indicates the best model should be ΔILD = 0.

In addition to aspect ratio, the aspect ratio mask can also be used to model the effect of the ratio of perimeter to area on ILD thickness variation using the same techniques mentioned above in connection to aspect ratio modeling. The resulting perimeter to area model and data as well as the aspect ratio model and data can be seen in Figure 5.10 for represen-
tative wafers from each pad split. Again, no statistically significant difference was seen across the pad split nor was any sizable perimeter/area dependence detected.

Finally, Figure 5.11 shows ILD thickness versus separation distance for the aspect ratio mask. Six different separation distance from 10mm to 60mm were used (see Section 5.3). In Figure 5.11, the density dependence was removed and all data points have the same aspect ratio. As the data shows, no systematic dependence on separation can be seen.

**Figure 5.10:** A comparison of the dependence of ILD thickness on perimeter/area (a) versus aspect ratio (b) for sample wafers from each pad type. The effect of pattern density has been removed to show only the perimeter/area or aspect ratio dependence.
From the models and confidence intervals displayed in Table 5.2, the significance of differences between pads across different models can be identified. In this experiment we find that the pattern dependent behavior of the two pads is very nearly the same except for a difference in pad interaction distance (3.4mm for the IC-1400 compared to 4.2mm for the IC-2000).

**5.7 Summary**

The masks and associated metrology and analysis methods presented in this paper represent important tools for rapidly evaluating the role of CMP process, consumable, and tool options on polishing performance, and should have application to other hybrid or

---

**Figure 5.11:** ILD thickness versus separation distance for the aspect ratio mask. Each data point represents a structure with the same aspect ratio but different separation distance. The density dependence has been removed to isolate structure separation effects.
novel planarization schemes. The data and models generated using the test masks can serve as an important base for evaluating and verifying physically based models and in investigating manufacturability/yield issues associated with CMP -- especially for process integration, development, or implementation issues. Finally, the resulting characterization models for ILD polishing as a function of key layout pattern dependencies have the potential for use in layout-specific evaluation of circuit performance degradation due to thickness variation.

In addition to contributing new mask designs for exploration of pattern-dependencies, several metrology tools and techniques have been presented, as well as analysis tools and methods, for generating ILD thickness models using relatively simple procedures. A simple pad comparison study experiment has been described. For the process and pads studied in this experiment, we find that density is the dominant layout factor contributing to pattern-dependent variation, and that area, pitch, and aspect ratio represent a second-order or no measurable effect.

Several extensions and applications of this work can be identified. The masks and analysis procedures detailed in this paper are being applied to extensive pad comparison studies and experiments to understand at both an empirical and physical level the role of conditioning and pad properties on ILD thickness variation. Investigation is also underway to evaluate novel process techniques such as silicon nitride capping layers for reducing intra-die ILD thickness variation using the masks. Finally, models extracted from characterization mask data are in use to study the impact of intra-die ILD thickness variation on circuit performance for a number of representative circuit architectures (e.g. SRAM arrays, standard logic cells) through modified layout parasitic extraction procedures.
Chapter 6

Variation Impact: Circuit Performance

6.1 The Nature of the Problem
This chapter discusses a methodology for determining the impact of pattern dependent or spatial variation on circuit performance. In the literature, significant research has been focused on understanding the impact of process variation on circuit impact [64-75], but the majority of this research assumes process variation to be completely random and drawn from either normal distributions or correlated normal distributions (i.e. Monte-Carlo type simulations or design centering methodologies). This approach is adequate when assessing the impact on lot-to-lot or wafer-to-wafer variability where the systematic nature of the variation is small compared to a larger normally distributed component. For spatial or pattern dependent variation, however, this approach is not completely acceptable because it essentially assumes that each location on a wafer or on a die has the same variance and mean inevitably leading to large over- or underestimates. Furthermore, this approach is especially unacceptable for interconnect variation since the effect of drawing from a random distribution ignores the fact that the variation is spatially correlated (i.e. two points located near each other on the same line are likely to have a smaller variance compared to points on the same line located a large distance away).

Currently, industry has been using “worst-case” based approaches or “guard-banding” to estimate the impact of variation on circuit approaches. In this approach, a technology file is built for several different process conditions (termed process corners) which are believed to capture the largest excursion the design is likely to see in the fab. For each technology file, the layout or circuit is extracted and simulated. “Worst-case” approaches are especially conservative in that they do not allow the designer to trade performance for
manufacturability and vice-versa in that it is impossible to know a priori which design violations are the most destructive and which are only slightly destructive in the sense of yield or reliability.

The methodology presented here represents a radical departure from statistical circuit simulation and guard banding. In our approach, the variation models (as developed and illustrated in the previous chapters) are used to explicitly model variation as a function of position on the wafer/die or as a function of the layout. This approach is deconstructive rather than aggregate in the sense that statistical circuit simulation relies on the central limit theorem of statistics to approximate the combination of many non-random elements

---

**Figure 6.1:** Originally statistical simulation techniques were used to understand the impact of process variation on circuit performance. This approach is successful and accurate for parameter variation which has a strong random component. For strong systematic components, the methodology presented here based on deterministic prediction of variation is more appropriate.

The methodology presented here represents a radical departure from statistical circuit simulation and guard banding. In our approach, the variation models (as developed and illustrated in the previous chapters) are used to explicitly model variation as a function of position on the wafer/die or as a function of the layout. This approach is deconstructive rather than aggregate in the sense that statistical circuit simulation relies on the central limit theorem of statistics to approximate the combination of many non-random elements.
of variation into one large random component. Our approaches obviates the need for estimating the mean and variance of a particular process parameter since each process parameter of interest is computed at every location on the die or wafer by leveraging existing variation models. Although each process parameter estimate has some variance and error (due to unmodelled effects or randomly driven events), this component is generally quite small. In actuality, the layout or process is generating variance rather than unknown and unattributable random causes. If substantial random and/or unexplained variance remains, this can be added to the systematic variation components using existing techniques.

![Diagram](image)

**Figure 6.2:** Estimation of the effect of systematic sources of parameter variation on circuit performance is hampered by a tendency of certain forms of systematic variation to be localized to several microns while other forms tends to assume more global dependencies which are outside the bounds of the electromagnetic region of interest in a simulation. In this way, TCAD types simulations alone are not sufficient.

The problem is especially vexing because of the intense coupling between design and manufacturing. A portion of the variance that a design sees is driven by the process (mainly lot-to-lot, wafer-to-wafer, and within-wafer), but an increasingly large part of the variance, especially for critical parameters such as the polysilicon critical dimension, is
dependent on the design itself as represented by the layout. In this way, the question “what is the impact of spatial or pattern dependent variation on my circuit?” cannot be answered in the general, but must be dealt with in the specific. As such, this chapter offers a methodology for determining the impact of pattern dependent or spatial variation on circuit performance which can be used on each design of interest or on several designs to gain insight rather than presenting a conclusive study or evaluation of the impact of spatial or pattern dependent variation on circuit performance. Several case studies will illustrate the methodology. Two case studies deal with interconnect variation and one case study explores pattern dependent device variation.

In addition to the inherent coupling between manufacturing and design, the general problem is especially difficult due to the nature of the variation. For example, ILD thickness variation tends to span large length scales while photo-lithography variation can exhibit both very local and more global length scales (see Figure 6.2). In this way, neither exhaustive TCAD-type simulations nor traditional ECAD (e.g. layout parasitic extraction engines or circuit simulators) alone can be employed to generate an effective methodology. It is the novel combination of elements from TCAD type simulators, ECAD type simulators, and custom tools which enables the methodology presented here.

Although considerable simulation time and accurate variation models are required to analyze a design in a circuit impact analysis, the result is valuable since it moves the design process closer to first pass prediction.

6.2 Variation Prediction

Although there are many process parameters which are subject to manufacturing variance, this work focuses on modeling only two factors: linewidth variation from photolithography and ILD thickness variation from CMP. Although other process parameters may also
exhibit pattern dependencies (e.g. microloading in plasma etch), these two are the most pattern dependent in present technologies and can directly modulate key process variables such as drive current or load capacitance.

6.2.1 Photolithography Modeling
Photolithography variation is difficult to model accurately since many physical factors are at work at different length scales. The work presented here on lithography variation modeling is an extension of the work presented by Yu [17, 76]. The total lithography error budget is broken down into smaller more manageable components. The following factors and percent contribution of each factor to the total die-level lithography variance can be identified:

<table>
<thead>
<tr>
<th>Component</th>
<th>% Lithography Variance</th>
<th>Spatial Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Optical Proximity</td>
<td>30%</td>
<td>1-3μm</td>
</tr>
<tr>
<td>2. Resist Effects</td>
<td>5%</td>
<td>&lt; 1μm</td>
</tr>
<tr>
<td>3. Depth of Focus</td>
<td>5%</td>
<td>3-4mm</td>
</tr>
<tr>
<td>4. Stepper Focus/Leveling Errors</td>
<td>20%</td>
<td>5mm</td>
</tr>
<tr>
<td>5. Reticle Errors</td>
<td>20%</td>
<td>1-3μm</td>
</tr>
<tr>
<td>6. Lens Aberration</td>
<td>15%</td>
<td>1-2mm</td>
</tr>
<tr>
<td>7. Random/Unknown</td>
<td>5%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 6.1: Several different components of systematic and random lithography variance and the estimated contribution of each factor to the total lithography variance [76].

The values in Table 6.1 are loosely derived from [76] and from industry experience.

Optical proximity components are estimated using aerial imaging [77, 78]. Aerial imaging has been in practice for approximately ten years and most simulators use the Hopkins model approach [78]. In this approach (see Figure 6.3), the intensity at the target $I(x,y)$ can be expressed as a non-linear function of the illumination lens and pupil, $J_0(x,y)$,
the projection optics $K(x,y)$, and the mask being imaged $f(x,y)$. The numeric aperture of the lens (NA), the illumination wavelength ($\lambda$), and the coherence of the imaging system ($\sigma$) are the key input parameters which shape the Hopkins model. Using these parameters and several assumptions of the image/projection system [77] allow the Hopkins integral to be expressed as a two dimensional convolution integral which allows the expression to be evaluated using numerically efficient Fast Fourier Transforms. By taking equi-illumin-

**Figure 6.3:** The Hopkins model used in aerial imaging simulators. Figure and notation derived from [77].
tion contours at the target, an estimate of the optical proximity component of lithography variance is enabled. A sample result is shown in Figure 6.4. The “ideal” features are shown as well as the results from aerially imaging.

![Figure 6.4: The “ideal” as-drawn features from an SRAM cell and aerial imaged features](image)

The depth of focus contribution is estimated by assuming that the amount of contributed variation is directly proportional to the amount of layer thickness variation or the non-planarity of the target. If the component due to depth of focus is represented as $\Delta_{CD}$, then

$$\Delta_{CD} = \alpha \times \Delta z(x, y)$$  \hspace{1cm} (6.1)

where $\Delta z(x, y)$ is the non-planarity of the underlying film and $\alpha$ is the proportionality constant which is selected to give a swing of 5% of the total lithography budget for the maximum swing in non-planarity of the target. A linear model is justified since even though the de-focus versus CD curve is highly non-linear the total excursion from depth of focus due to non-planarity of the target is a fraction (typically 25-50%) of the total depth-of-focus window where linear behavior is more typical. The amount of non-planarity of the target film is estimated by estimating the ILD thickness variation (discussed in the previous chapter and in the next section) and propagating this variation up to ensuing layers.
The stepper/leveling errors are unique among the other terms in lithography in that they are systematic within the field but random from field to field. In this way, this source is neither random nor systematic but hybrid. Estimating this factor begins by assuming that the within-field variation is best approximated by a linear slanted plane:

$$\Delta_{\text{STEPPER}} = \beta_0 x + \beta_1 y$$  \hspace{1cm} (6.2)

where \(x\) and \(y\) are spatial coordinates within the field and \(\beta_0\) and \(\beta_1\) are normal random variables which vary from field to field. They are zero mean variables with \(2\sigma\) set to half the total stepper variability.

The lens aberration component is difficult to predict since every stepper system has a slightly different profile. From industry observations and based on data observations in [76], to first order the aberration component can be represented as a term proportional to the square of the radius from the center of the stepper field:

$$\Delta_{\text{LENS}} = \gamma_1 r^2 + \gamma_2$$  \hspace{1cm} (6.3)

where \(r\) is the radial distance from the center of the stepper field and \(\gamma_1\) and \(\gamma_2\) are adjusted so that \(\Delta_{\text{LENS}}\) is 15% of the maximum lithography variance.

The remaining terms are the most difficult to estimate. For this reason and since they are localized to very small distances (and thus unlikely to correlate over a long net and impact performance), they are left unmodelled. If the reticle of the design was measured, then the reticle component can be directly determined (via Yu’s method in [76]). For resist modeling, several results in the recent literature indicate that the amount of cross-linking can be estimated by looking at the slope of the intensity function \(I(x,y)\) at the boundary of all features.
6.2.2  ILD Thickness Variation

The ILD thickness variation model for CMP processes discussed in Chapter 5 and in [59] is used to estimate the ILD thickness variation for each layer:

\[ z = z_0 - z_1 - K t + \rho z_1 \quad \text{for} \quad t > \frac{z_1}{K}. \quad (6.4) \]

The pattern density component (\( \rho \)) for the optimized interaction distance (using the modeling methods discussed in Chapter 5) is estimated rapidly by realizing that the area within a bounding region (the density window) for all points on a layout can be phrased as a convolution of the layout database, \( f(x,y) \), with the density window function \( w(x,y) \), in this case a square region of size equal to the interaction distance. Hence, this convolution can be computed rapidly using 2-D FFTs. Careful consideration, however, is needed where the density window does not completely overlap the layout region (i.e. for regions near the edge of the layout). Two solutions are recommended. The first option circumvents the problem by replicating the layout around the edges of the layout. The second approach uses a cyclical 2D-FFT which maps coordinates which fall of the left edge of the layout to the right side of the layout and vice-versa (i.e. cyclical boundary conditions). The first approach is less desirable, however, since it consumes much memory compared to the second method where only one copy of the layout is needed. Also, the first approach may require smashing the hierarchy while the second method can still be computed in a hierarchical or pseudo-hierarchical (i.e. flatten as you go) approach.

6.3  Experimental Methodology

Figure 6.5 shows the methodology used to estimate the effect of device and interconnect systematic variation on circuit performance. Using the appropriate variation models, the input layout, or “ideal” or “as-drawn” layout, is transformed into the “as-manufactured” layout. In addition, tables are built for layer thickness variation as a function
of position on the layout for all layers of interest. Next, devices and resistors are extracted from the “as-manufactured” layout using traditional layout-parasitic extractors (LPE) such as Dracula [80]. The connectivity of the interconnect layers is then established, and the networks of interest (or all of the networks) are selected and segregated from the original layout. In addition, all features within a specified distance from the nets of interest (the “halo”) is segregated. The nets of interest and halo are meshed and capacitance values are computed using either analytical expressions when the simplicity of the structure allows or full 3-D solvers for more complicated geometries. Finally, the capacitance matrix is combined with the extracted resistance and device data to form a net-list suitable for simulation in circuit simulators such as SPICE. By turning individual sources of variation on or off (such as lithography and ILD thickness variation), the impact of individual sources of variation on circuit performance can be determined. These stages are discussed in more detail in the following sections.
Figure 6.5: The methodology for estimating the impact of systematic pattern dependent variation on circuit performance. In this methodology, key elemental steps include meshers, layout-parasitic extraction engines, 3-D solvers or analytical capacitance equations, and variation models.
6.3.1 Connectivity and Device and Resistance Extraction

Device and resistance extraction are easily done using traditional layout parasitic extraction tools. These tools convert layout polygons into device and resistance components through boolean operations (e.g. the gate region of an MOS transistor can be defined as the logical “and” of poly with active). In addition to extracting raw device and resistance components, the location of the transistors and resistors in x,y coordinates is also recorded. This information is used at the tail end of the flow shown in Figure 6.6 to determine which resistors and devices belong to the nets of interest. Since interconnect lines can run long lengths especially for higher levels of metal, long lines need to be broken into smaller sections. In this way, a resistor can be localized to a region within a layout rather than lumped over the entire length of the line.

In addition to extracting devices and resistors, electrical connectivity is needed. Electrical connectivity is usually a network graph (see Figure 6.6) which depicts which polygons are physically connected to other polygons with some information usually attached to the graph weight such as the distance between the polygons. In addition, all networks are named with a unique identifier number or with the text associated placed on the network (referred to as texting).
### 6.3.2 Net and Halo Segregation

After stamping all networks on a layout with connectivity, the critical nets of interest (i.e. all nets which the designer believes are most critical to the speed and operation of the circuit) or all nets (if the circuit is small enough) are segregated from the layout as shown in Figure 6.7. Segregating is the process of isolating the nets from the rest of the layout. The isolated nets are usually deposited in a separate file or cell instance. In addition, all polygons within a specified distance of the nets of interest are selected. This is called segregating the net “halo.” The distance to use for the halo is selected to be greater than the electromagnetic coupling present between the nets of interest and the surrounding environment. In practice, this distance is often on the order of 10 microns. All nets in the “halo” are assumed grounded. In this way, the designer will be able to calculate not only the self capacitance of a net and the coupling capacitance to neighboring nets, but also the capacitance to the neighborhood. The primary purpose of segregation is to allow the criti-
cal nets capacitance to be estimated using full 3-D solvers. The entire layout cannot be fed into a 3-D solver because of finite simulation time, and 1-D and 2-D LPE extraction engines cannot be used because of gross inaccuracies.

6.3.3 Gridding and Meshing

After segregating the nets and halo from the layout, the nets and halo need to be gridded into small regions. The gridding is necessary for use in 3-D field solvers. In our methodology, both rectilinear and Delaunay gridding schemes are employed. Rectilinear gridding is faster, but cannot be used on aerial imaged geometries because of rounded corners and sharp edges. For these features, Delaunay gridding is used in which the top and bottom sides of all regions is gridded with equilateral triangles. The side faces of all features are gridded with quadrangle faces. An example of rectilinear gridding and Delaunay gridding is shown in Figure 6.8.

Regardless of the gridding system used, often the total number of panels created during the gridding session is larger than the capacity of the 3-D solver. In this case, two solutions are possible. First, the number of panels can be reduced by forming panels with a larger area. This method works, but compromises the accuracy of field simulation. The second technique involves forming simulation windows. In this procedure, the gridded regions are broken up into overlapping regions (see Figure 6.8b). All the panels in each

Figure 6.7: The nets of interest as well as the net “halo” are segregated from the layout. In this way, the critical nets of interest can be simulated in 3-D solvers.
field simulation region are simulated in a separate field simulation. The simulation windows must overlap because the coupling between panels near the edge of a simulation window to neighboring panels needs to be computed. The contribution of the capacitance from the overlapping regions, however, must not be double counted to preserve simulation integrity. Double counting can be avoided by looking at the panel-by-panel contribution to the charge on each net and always assign the charge to only one region. The size of the overlap is usually set on the order of a few microns compared to a typical window size of 100 microns by 100 microns.

6.3.4 Capacitance Extraction and Circuit Simulation
After gridding the critical nets and halo into panels over several simulation windows, a 3-D field solver is used to generate the capacitance matrix for the nets and halo of interest. In all cases, the halo regions are left grounded and average case coupling capacitance values are chosen (rather than worst case 2x rules) since information about activity factors is not available.

A sample capacitance matrix for two nets plus the halo region is:
\[ C = \begin{bmatrix}
    C_{11} & C_{12} & C_{1H} \\
    C_{21} & C_{22} & C_{2H} \\
    C_{H1} & C_{H2} & C_{HH}
\end{bmatrix} \]  

(6.5)

where \( C_{12} \) is the capacitance from net “1” to net “2”, \( C_{1H} \) is the capacitance from net “1” to the halo, and \( C_{2H} \) is the capacitance from net “2” to the halo. Because of symmetry, \( C_{12} = C_{21} \), and \( C_{1H} = C_{H1} \). The capacitance from net “1” to ground (\( C_{1\text{-GND}} \)) is computed using

\[ C_{1\text{-GND}} = C_{11} - C_{12} - C_{1H}. \]

Analogous formulas hold for the capacitance from net “2” to ground. The process repeats for each simulation window.

After computing the capacitance matrices, the resistance information needs to be combined to form a lumped R-C network. In our methodology, a \( \pi \) ladder network model is used. In this model the number of resistors in each simulation window is counted on a net by net basis and scaled by a factor of two (referred to as the node count number). The capacitances to halo and ground in each simulation window are then summed (since this represents the capacitance to electrical ground) and divided by the node count number. These values are then distributed across all the resistor connections within each simulation window. The coupling capacitance between two nets is instantiated by placing a capacitor from the centroid of the first net in the simulation window to the centroid of the second net in the simulation window. If the centroid of the net does not lie physically on the net, the closest location on the net to the centroid is selected. An illustrative figure for the case of two nets is shown in Figure 6.9. Because the coordinates of each resistor was stored during LPE extraction, it is known with considerable accuracy which resistors belong to which simulation window.

In some special cases where the geometries and associated halo are comparatively simple (e.g. buses crossing other buses in a regular fashion or relatively isolated lines over
a ground plane), gridding and 3-D solving may not be necessary. In this case, simple numerical analytical equations can be used (e.g. [81]). This can yield extremely rapid simulation results, but is only valid for very regular geometries.

\[
CA = \frac{(C_{P1-GND} + C_{P1-HALO})}{2n_1}
\]

\[
CB = \frac{(C_{P2-GND} + C_{P2-HALO})}{2n_2}
\]

\[
CC = (C_{P2-P1})
\]

Figure 6.9: This figure shows a sample π ladder distribution scheme for distributing resistors and capacitors within a simulation window. In this figure, the node count numbers \((n_1, n_2)\) are 6 and 3. In this figure, the “halo” geometries are not shown.

The resulting netlist coupled with the devices can then be analyzed as is (i.e. as the ratio of different capacitance terms with variation modeling turned on versus off) or the netlist can be fed into a circuit simulator (e.g. SPICE) and delays extracted. The ratio of the delay with variation sources turned on to the delay through a critical net without varia-
tion (i.e., with ideal geometries) is a useful metric for comparing the impact of systematic pattern dependent variation on circuit performance. Also, by making the same comparison with different variation sources turned off, a pareto chart assessing the impact of each source of variation from greatest to least impact on circuit performance is obtainable.

6.4 Case Studies
In this section, three different case studies are examined as an application of the methodology. Two case studies relate to interconnect variation and the third study is focused on the impact of device variation. For the interconnect variation examples, one study uses the analytical equations to compute capacitance values while the second case study relies on gridding and 3-D solvers.

6.4.1 A Balanced Clock Tree
Figure 6.10 shows the layout of a simple balanced clock tree [82]. Several large embedded memory and logic blocks have been placed on the layer below the clock tree. Each block is assumed to have relatively constant density and was extracted from other industrial layouts. The length of all paths is the same from the clock source to each end of the tree (denoted path 1, 2, 3, 4 in Figure 6.10), and all branches are half as wide as the previous branch for resistance matching [82]. Also, a local ground plane has been inserted under the clock tree (Figure 6.10 inset); thus, there should be no designed skew between any of the paths in the clock tree. However, since the clock tree runs over different densities, there will be ILD thickness variation along each path which will lead to capacitance variation and clock skew. The estimated ILD thickness variation is shown in Figure 6.11. Figure 6.12(a) shows the estimated ILD thickness variation for two of the paths shown in Figure 6.10 as well as the ideal case (ILD thickness uniform at 1μm). The ILD thickness variation model discussed in Chapter 5 and presented in [59] was used and an interaction
distance of 4mm was used. The metal thickness for the clock line was set at 1.5 μm while the underlying structures and ground plane had a metal thickness of 1.25μm.

**Figure 6.10:** A simplified balanced H-bar clock tree. Since the length of the line is the same from the buffer (at the center of the layout) to the end of each path and a ground plane runs under the clock line, the skew should ideally be zero. However, the pattern density underlying the clock line changes, and the realistic clock performance will skew to some degree constrained by the amount of ILD thickness variation.

Metal-CD variation can be safely ignored due to the large width of each path (100 microns in the center clock structures and 50 microns in the second stage). Figure 6.12(b,c) shows the linewidth and resistance as a function of position along the net. Figure 6.12(d) is the estimated capacitance versus position along the net. The capacitance was extracted using analytical formulas [81] for the capacitance from an isolated line to a ground plane. For computing the resistance and capacitance values, a simulation window
size of 100 microns was used. In this way, all capacitance and resistance values in Figure 6.12 are reported as units per 100 microns of line length, and lumped accordingly.

Figure 6.11: The estimated ILD thickness variation for the balanced H-bar clock tree in Figure 6.10. The ILD thickness variation model in [59] was used to estimate the amount of ILD thickness variation. If the ILD thickness was completely uniform, a balanced H-bar clock tree should be skew free; however, ILD thickness variation perturbs the balance of the tree resulting in some clock skew.

Figure 6.12(e) shows the simulated voltage wave forms for two paths as well as the ideal path. As shown, a 7% skew was observed between the two paths and a 17% skew was observed between path (3) and the ideal case. In each case, the skew is reported as the ratio of the difference between the delays of two paths divided by the delay of the “ideal” path (i.e. with the ILD thickness held constant at one micron across the entire layout).
6.4.2 A 256K SRAM Array
Figure 6.13 shows a 256Kx1 SRAM array for use in a 0.5\(\mu\)m process. The layout is approximately 7mm x 8mm in size, and the array is made up of sixteen 16K SRAM
macro-cells with columns and drivers along the center of the layout. A ring of probe pads encapsulates the SRAM array. Bit lines run on metal-2 from the edges of each 16K cell to the decoders giving a total run-length of approximately 3mm for the bit lines. Power and ground lines are fed horizontally on metal-3 while local connections are placed on metal-1.

![Diagram of SRAM array](image)

**Figure 6.13:** A 256Kx1 SRAM array implemented in a 0.50 micron three level metal process. The array is constructed from sixteen 16K macro-cells plus decoder circuitry along the central region.

The estimated ILD-1 and ILD-2 thickness variation is shown in Figure 6.14. The same model was used to predict ILD thickness variation as in the previous case study. Note the drop off moving from the center of the layout to the edge of the array along the x direction. This drop off is caused by the interaction of the SRAM array with the probe pads. The
interaction is not present in the y direction because probe pads were only placed on the left and right sides of the layout.

![ILD-1 and ILD-2 Thickness Variation](image)

**Figure 6.14:** The ILD-1 (between metal-1 and metal-2) and ILD-2 (between metal-2 and metal-3) thickness variation for the 256Kx1 SRAM array shown in Figure 6.13. The total range is approximately 0.2μm in each case.

Figure 6.15 shows the model used to compute the metal-CD variation as a function of position. As discussed in section 6.2.1, a decomposed model consisting of depth of focus, stepper leveling, optical proximity, and lens aberration effects is used. Aerial image simulation was used to estimate optical effects but did not contribute significant metal-CD variance to this 0.5μm process. For smaller geometries (e.g. for 0.25μm and below processes, the variance component from optical proximity effects can be substantial although local-
The coefficients for each variance components of systematic lithography error is also shown in Figure 6.15.

\[ \Delta CD = (\alpha \Delta ILD) + (\beta_1 x + \beta_2 y) + \left( \gamma_1 r^2 + \gamma_2 \right) \]

**Figure 6.15:** The estimated metal-CD variation for one of the metal layers in Figure 6.13. Since the stepper leveling contribution is random from field to field, this figure represents only one of a family of delta-CD contours that are possible.

This study is specifically focused on variation in bit-line capacitance (which translates to read access time skew) from location to location in the array. Figure 6.16 shows a 100μm x 10μm slice from a bit-line near the center of the array. The figure shows the result of segregating the bit-line and the halo and meshing the structure using Delaunay gridding. Since the array structure is so similar from bit-line to bit-line, this structure can
be used for all bit-lines in the array with appropriate adjustment of dielectric thickness. The length of the bit-line is also longer than 100µm, but again, the periodicity of the array allows replication of the structure vertically to achieve a representation for nearly the entire length of a bit-line (approximately 3mm).

![Diagram](image.png)

**Figure 6.16:** A 100µm by 10µm slice from the 256Kx1 SRAM array showing the bit-line surrounded by the halo structures above, below, and to either side. Structures below Metal-1 (such as poly and active area) did not couple significantly to the bit-line structures; thus, they were removed.

In the photolithography model discussed in Section 6.2.1, the stepper contribution is systematic within field and random from field to field. In this way, a number of different simulations need to be run to simulate the natural field-to-field variability present in stepper leveling variance. Instead of simulating many Monte-Carlo runs around the gridding, capacitance extraction, resistance extraction, and circuit simulation (which would take an excessive amount of time), a Design of Experiments (DOE)/macromodelling approach is useful. In this approach, a DOE is generated for several of the key physical parameters for the bit-line structure. These factors, the number of levels, and the design range are shown
in Table 6.2. The full simulation methodology (Figure 6.5) is executed for each run. Ini-

<table>
<thead>
<tr>
<th>Factor</th>
<th>Number of Levels</th>
<th>Level Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ Metal-1 linewidth</td>
<td>5</td>
<td>-0.1μm -&gt; 0.1μm in equal steps</td>
</tr>
<tr>
<td>Δ Metal-2 linewidth</td>
<td>5</td>
<td>-0.1μm -&gt; 0.1μm in equal steps</td>
</tr>
<tr>
<td>Δ Metal-3 linewidth</td>
<td>5</td>
<td>-0.1μm -&gt; 0.1μm in equal steps</td>
</tr>
<tr>
<td>Δ ILD-1 thickness</td>
<td>5</td>
<td>-0.1μm -&gt; 0.1μm in equal steps</td>
</tr>
<tr>
<td>Δ ILD-2 thickness</td>
<td>5</td>
<td>-0.1μm -&gt; 0.1μm in equal steps</td>
</tr>
</tbody>
</table>

Table 6.2: The DOE experiment used to construct a macro-model of bit-line capacitance versus metal linewidth and ILD thickness values.

tially, however, a two level with center point full-factorial DOE was run on the structure to look for the presence and strength of interaction terms. This study revealed that only the main effects were playing a significant role in determining the capacitance from the bit-line to ground and the halo. Thus, only a main effect DOE had to be run on the experiment in Table 6.2. Figure 6.17 shows the results of this experiment. In addition, a macromodel of the bit-line capacitance to the five physical parameters was fit.
With the macro-model, a rapid conversion between physical parameters and bit-line capacitance is enabled. Using the ILD thickness variation data (Figure 6.14) and Monte-carlo runs on Δ metal CD for each layer, a distribution of bit-line capacitance across all bit-lines on the layout can be generated. Figure 6.18 shows an example. This plot clearly reveals significant skew in capacitance for bit-lines near the edge of the array versus the

\[
C = \frac{4.74\Delta/l_{w1} + 14.32\Delta/l_{w2} + 1.2\Delta/l_{w3} + 19.06}{ILD_1^{0.19} \times ILD_2^{0.14}}
\]

**Figure 6.17:** A DOE and macro-model was generated of the bit-line capacitance to ground and halo (i.e. with the halo grounded). This model allows for a rapid conversion between variation in these five physical parameters to bit-line capacitance.

With the macro-model, a rapid conversion between physical parameters and bit-line capacitance is enabled. Using the ILD thickness variation data (Figure 6.14) and Monte-carlo runs on Δ metal CD for each layer, a distribution of bit-line capacitance across all bit-lines on the layout can be generated. Figure 6.18 shows an example. This plot clearly reveals significant skew in capacitance for bit-lines near the edge of the array versus the
center. Also, the center bit-lines show more variance compared to edge bit-lines. Naturally, this result is only as good as the metal-CD and ILD thickness variation models, and should change from process to process and design to design.

**Figure 6.18:** The capacitance of a bit-line near the edge and near the center to ground and the halo structure (held grounded) for the 256Kx1 SRAM array. Significant skew can be seen for bit-lines near the edge versus bit-lines near the center. Also, bit-lines near the edge show more variance compared to near the center.
6.4.3 Poly-CD Variation for a 64x8 SRAM Cell

As opposed to the previous two case studies which focused more on systematic interconnect variation, this case study focuses on systematic device variation, particularly poly-CD variation. Figure 6.19 shows the layout of a 64x8 SRAM macrocell. The layout is approximately 180\(\mu\)m x 140\(\mu\)m in size and contains approximately 4000 transistors. The layout was originally designed for 3\(\mu\)m design rules but was scaled down to 0.25\(\mu\)m design rules.

![Figure 6.19: A 64x8 SRAM macro-cell. The layout is approximately 180\(\mu\)m x 140\(\mu\)m](image)

Only aerial imaging was used to simulate the within-die poly-CD variation. This is mainly due to the small size of the layout compared to the length scales of other sources of variation (e.g. stepper leveling errors, lens aberration). As before, resist and reticle variation modeling were not done because of the lack of good models. Figure 6.20 shows a surface plot of \(\Delta\) poly-CD variation versus location on the cell. Note the presence of periodic variation...
components in addition to larger peaks and valleys. Interconnect variation modeling was not done due to the small size of the layout and was only crudely modeled using identical load capacitors at each output bit. Since no interconnect variation modeling was used, the meshing, capacitance, and resistance extraction steps could be skipped.

Simulated Die (Cell) Level Variation

Figure 6.20: The estimated poly-CD variation across all coordinates in the cell. Aerial imaging simulation was used to estimate the poly-CD variation.

For accurate circuit simulation, 0.25μm SPICE models were needed, but could not be obtained from industrial connections as they were held confidential. Instead, a 0.25μm device model was created based on SIA roadmap specifications [83,84] and typical values reported in the literature. The sensitivity of $I_{DSAT}$, the measured drain current when both the gate and drain are held at the supply rail, is shown in Figure 6.21. Although the $I_{DSAT}$
values are slightly higher than normally expected, the sensitivity to channel length variation is similar to results published elsewhere (e.g. [85]).

![Graph](image1)

**Figure 6.21:** (a) Drain current versus drain voltage for $V_{gs}$ at the supply rail. (b) Saturation or drive current versus channel length for the n-mos and p-mos devices.

Figure 6.22 shows the result of simulating the aerially imaged layout with the above device models. The voltage at each output line is shown for the address set to 00000 and for strobing the input data lines from 11111111 to 00000000. Also, the write strobe is held high resulting in a display of write-access behavior. In this case study, we are interested in the time required to write a data word to an address location. In particular, the maximum skew for all the data lines is of key interest since this is a measure of the susceptibility of this circuit to poly-CD variation. Since interconnect variation has only been crudely approximated, the amount of skew for the “as-drawn” layout is near zero, but as Figure
6.22 shows, there is significant skew for the write access time. The read access time does not show significant skew behavior; thus, it is not discussed.

The behavior shown in Figure 6.22 is just for one particular address location. As Figure 6.23 shows, the maximum amount of write-access time skew is heavily dependent on the input address vector. Oscillating behavior is especially apparent on address line Out[5] which is most likely caused by so-called odd/even strobing. The 64x8 SRAM cell is composed of two 32x8 cells. The odd address lines are routed to one 32x8 cell while the even lines are routed to the other 32x8 cell. This causes odd and even address to behave differently especially if there is a large device variation component. This result implies that the results of simulating the impact of spatial variation on circuit performance can be heavily dependent on the simulating conditions used. Thus, careful planning and examination of test vectors and simulating conditions is needed.
In addition to address dependent behavior, this layout also shows simultaneous switching phenomenon. This behavior occurs when the input address and input data switch at
nearly the same time instant. Figure 6.24 shows the result of simulating this condition. The amount of skew has now increased tremendously to almost 60% of the worst-case delay. Again, this result clearly demonstrates the need to carefully determine the simulation conditions before interpreting results.

6.5 Discussion

The case studies presented in this chapter seem to indicate that systematic interconnect variation impacts circuit performance, but only as a second-order effect while systematic device variation can substantially impact circuit performance. This finding is also intuitive considering that critical interconnect lines are often very long. Over the entire length of the line, linewidth and ILD thickness parameter can change significantly, but the law of large numbers tends to force the electrical behavior of the network to depend on the mean physical values as opposed to variance. For device variation, however, the drive action of a
critical net is extremely localized to only several small devices. If even a small change occurs at one of these critical devices, then the entire electrical behavior of the critical net is altered rather than dampened from averaging.

Systematic interconnect variation, however, cannot be ignored for three important reasons. First, if a short interconnect plays an important role in determining the speed or functionality of a design, then systematic interconnect variations will not be damped out by averaging due to the spatial locality of the network. Second, in certain circumstances (such as the clock tree example), an interconnect parameter may shift over the entire or a large portion of the length of a network. In this case, the entire mean of the network has changed, and thus, the delay characteristics of the net. Finally, within-wafer variation can substantially impact circuit performance since variation of this type ultimately results in shifting the entire mean of the parameters in a design up or down depending on the location on a wafer.

The results presented in the case studies seem to indicate that the primary metric for diagnosing the impact of systematic variation on a network is the degree to which variation is correlated over the length of the network. If variation in a key physical parameter is correlated over a long distance then this essentially translates into a shift in the overall mean in the physical characteristics of the network and ultimately the electrical behavior. Short correlation distances lead to an averaging of systematic variation and a dampening of the impact of variation on circuit performance. Device parameters are localized within a micron; thus, any amount of variation in the device will directly translate into a variance in the electrical behavior of the network. If all of the device variance along the path of a critical net is highly correlated, the effect will be accentuated due to a complete shift in most or all of the transistor in the path of the critical net. Less correlation, although still capable of impacting electrical behavior, dampens the impact on electrical behavior.
Chapter 7

Variation Reduction

7.1 Introduction
The previous chapter explored the impact of systematic variation on circuit performance; the focus of this chapter is on determining methods for reducing variation. Variation reduction is worthwhile since a large variance or an out-of-control tool can be a yield loss mechanism. In particular, this chapter deals with reducing ILD thickness variation in CMP processes through the use of metal-fill patterning. Improving uniformity in a CMP process can reduce slurry usage and increase pad lifetime ultimately leading to lower cost of ownership. Also, as the die size has grown larger, the amount of ILD thickness uniformity has decreased as illustrated in Figure 7.1.

Attempts to control CMP intra-level dielectric thickness variation include an exhaustive search for and experimentation with different consumable and process choices (especially pads), but no consumable choice currently available appears to reduce appreciably pattern-dependent dielectric thickness variation [37]; thus, the only viable choice available for reducing layout pattern dependent dielectric thickness variation is to change the layout pattern itself via the introduction of metal-fill patterning. Metal-fill patterning is the process of filling the large open areas on each metal layer with a metal pattern, which is either grounded or left floating, to compensate for pattern-driven variations.
Note that metal-fill patterning practices are an "intrinsic" integration issue, i.e. the problem cannot be solved either at the unit process step or as a circuit design issue alone; there is a need to integrate process and design concerns and deal with the problem as a whole. Improvements in uniformity at the process/CMP module level resulting from metal-fill patterning practices must be carefully checked against design/electrical concerns of any added interconnect capacitance resulting from metal-fill.

**Figure 7.1:** Range of ILD thickness variation versus die size. As the die size has increased, the range of ILD thickness variation has also increased. This can be attributed to the size of the chip approaching and then surpassing the finite planarization length of the CMP process. The data shown was simulated using the model in [89]. The same layout was used for each data point and scaled appropriately. Although this would seem to indicate that scaling should help the problem, in reality, scaling forces are checked by ever increasing demands for higher transistor counts leading to larger and larger die sizes.

Die Size (mm$^2$) vs. ILD Thickness (μm)
Because of the confidential nature of metal-fill patterning practices and design rules in general, relatively little information about metal-fill patterning practices has been publicly reported. Ichikawa, et al. [86] describe a metal-fill patterning practice for planarizing a five-level spin-on-glass (SOG) interconnect CMOS process. A procedure for automatically generating metal-fill patterns is presented and some consideration is given towards optimizing the metal-fill design-rule to meet a given planarity target and to reduce the effect of added interconnect capacitance associated with the metal-fill; however, this methodology was developed for SOG processes, the consideration of capacitance effects is not described completely, and simulation/modeling aspects related to capacitance effects are not rigorously explored or stated. Camilletti [88] described and explored a metal-fill patterning practice in a CMP process, and reported significant improvements in uniformity. Stine, et al. [52] also explored the effects of metal-fill patterning practices on dielectric thickness uniformity and presented a mechanism for the beneficial increases in uniformity via pattern density modeling. In both these papers, however, interconnect capacitance and design rule optimization concerns were not addressed.

This chapter describes a unified methodology for designing and optimizing metal-fill design rules and procedures which is suitable for inclusion in automated CAD tools and which deals with both CMP process/uniformity concerns at the module level and capacitance/electrical concerns at the design level. While we are concerned with methods to help support automatic generation of metal-fill, we are not concerned with the actual layout generation algorithms. Also, although the beneficial effects of metal-fill patterning practices will be reviewed and the mechanisms for this improvement described fully, we are particularly interested in developing a link between specific metal-fill design rules and the
resultant improvements in ILD thickness uniformity. Finally, we are especially interested in the key integration issues associated with optimizing metal-fill design rules to minimize the accompanying increase in interconnect capacitance. Although we only consider metal-fill patterning practices for traditional back-end-of-line interconnect processes, the generic methods presented here can potentially be adapted to shallow-trench, damascene, and other inlaid polishing processes.

7.2 The Effects of Metal-Fill on Dielectric Thickness Uniformity

In this section, we demonstrate via two industrial-based experiments the beneficial effects of metal-fill patterning practices on dielectric thickness uniformity. The first experiment was conducted on existing test vehicles as an initial feasibility and proof-of-concept study. The second experiment was conducted on an actual product and used a more aggressive patterning practice in an effort to explore more fully the gains possible using metal-fill patterning practices.

In the first experiment, we used a standard test vehicle containing SRAM, defect density test structures, and device arrays. Two versions of this mask set were produced. The first mask set, reticle “A”, did not contain any metal-fill structures while the second mask set, reticle “B,” contained metal-fill structures. The fill pattern used in this experiment was used as a buffer to fill these large open areas greater than 250μm x 250μm between adjacent test structures and circuitry. Also, no metal-fill pattern was placed less than 40μm away from any active circuitry. The metal-fill implementation on reticle “B” also needed structures to accommodate yield inspection, electromigration testing, capacitance considerations, and device characterization needs. Thus, only a portion of the total eligible area for reticle “B” received metal-fill with the metal 2 level incorporating slightly more pat-
tern fill compared to the metal 1 level. A 0.35μm CMOS technology was used to fabricate the test structures. The ILD thickness was measured optically on nine die (approximately 17 sites per die) on each wafer for several lots to yield approximately 30,000 measurements. The site locations on each die were selected from populations near the thickest regions and near the thinnest regions.

Figure 7.2 shows the ILD thickness distributions (all thicknesses have been normalized) for one die for structures patterned with reticle “A” (Figure 7.2a-b) and reticle “B” (Figure 7.2c-d) at the ILD1 and ILD2 level. The standard deviation and uniformity for reticle “B” with pattern fill has improved by approximately 20-25% and by 15% at the ILD1 and ILD2 levels respectively compared with reticle “A” with no pattern fill. The results for ILD 2 are not as pronounced as for ILD1 because the density of underlying topography in ILD 2 before metal-fill was more uniform to begin with compared to the underlying topography in ILD 1 before metal-fill. As the ANOVA table for the ILD-1 level in Table 7.1 shows, the difference in standard deviation between reticles (i.e. with metal-fill and without metal-fill) was the only statistically significant difference observed (as opposed to wafer-to-wafer or die-to-die type variation). Similar results were observed at the ILD-2 level.
In the second experiment, the same experimental methodology and process technology was used: two versions of a reticle were generated, one without metal-fill (Reticle “A”) and one with metal-fill (Reticle “B”). In this case, however, a much more aggressive metal-fill patterning strategy was implemented and actual product layout at the metal-2 layer was used in the experiment instead of that for a test vehicle. The ILD thickness was measured optically. For intra-die measurements four sites/die (selected to achieve the largest range of variation) were measured on three dies per wafer across three wafers for each split. For within-wafer and wafer-to-wafer measurements one site on twenty-one dies per wafer was measured across all six wafers in each split. Substantial improvement was observed at the intra-die level (60% reduction) as well as within-wafer (35-40% reduction) and wafer-to-wafer (35-40% reduction). Intra-die variation for some products can be as high as 0.6µm total range and 0.15µm within-wafer [87]

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Sum of Sq.</th>
<th>Mean Sq.</th>
<th>F Value</th>
<th>Pr(F &lt; 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wafer</td>
<td>1</td>
<td>21681</td>
<td>21681.5</td>
<td>0.93363</td>
<td>0.336</td>
</tr>
<tr>
<td>reticle</td>
<td>1</td>
<td>314918</td>
<td>314917.9</td>
<td>13.56</td>
<td>0.000</td>
</tr>
<tr>
<td>die</td>
<td>1</td>
<td>44408</td>
<td>44407.6</td>
<td>1.91</td>
<td>0.169</td>
</tr>
<tr>
<td>wafer x reticle</td>
<td>1</td>
<td>19</td>
<td>19.5</td>
<td>0.00</td>
<td>0.977</td>
</tr>
<tr>
<td>wafer x die</td>
<td>1</td>
<td>43585</td>
<td>43584.7</td>
<td>1.88</td>
<td>0.173</td>
</tr>
</tbody>
</table>
In addition to the improvements in dielectric thickness uniformity, other benefits including improved etch and lithography process uniformity are often observed. Also, metal-fill patterning of test vehicles, such as in the first experiment, has the additional benefit of allowing the test vehicle to mimic more closely the behavior of real-life products and reduce product learning cycles [88].
7.3 Mechanisms for Metal-Fill Patterning Benefits

The improvements in dielectric thickness uniformity observed in metal-fill experiments can ultimately be attributed to layout pattern-density as discussed and shown in Chapter 5. According to (5.1), if a layout has a full scale range of pattern density before metal-fill and sufficient metal-fill is added to reduce the range in pattern density by 50% then the resulting dielectric thickness variation should be reduced by half. Furthermore, for a given CMP process adding metal-fill tends to increase the mean layout pattern density (and dielectric thickness) across the chip resulting in thicker oxide films and correspondingly lower layer-to-layer capacitance values than would be seen in a non metal-fill layout.¹

Design decisions about the size and extent of metal-fill patterning can be made based on models similar to (5.1). These decisions can be made by assuming a worst case scenario (such as a square open region free of metal 1mm by 1mm in size) and examining the effect of adding metal-fill. For example, if one considers a metal-fill patterning scheme of vertically oriented lines (see Figure 7.3 inset) with a buffer distance defined as the distance between the nearest active circuit region and a metal-fill line of 25μm, a design chart can be generated (e.g. Figure 7.3) showing the pattern density that can be achieved inside this 1mm x 1mm square region for a given linewidth and linespace of the metal-fill design. If it is desired to reduce the dielectric thickness variation in half, then a linewidth and linespace for the metal-fill design rule should be chosen which would give a pattern density inside the 1mm x 1mm square box of at least 50%. As Figure 7.3 shows, there are many possible design rules along the 50% contour which can meet this requirement.

¹ Although increasing the dielectric thickness has the effect of reducing layer-to-layer capacitance, coupling or line-to-line capacitance may actually increase with increasing dielectric thickness due to decreased shielding effects.
Choosing a value along this contour as well as other design rule issues (e.g. the choice of the buffer distance and whether to use lines or square blocks) are dictated by electrical/capacitance considerations and are dealt with in the next section.

![Contour Plot of Achievable Pattern Density](image)

Figure 7.3: A contour plot of the achievable pattern density inside of an empty 1mm x 1mm area versus the metal-fill design rule. The metal-fill patterning scheme is shown in the inset. The buffer distance for this example has been fixed at 25\(\mu\)m. A minimum linewidth of 0.35mm and a minimum linespace of 0.45\(\mu\)m has been assumed.

Although the above method of selecting a metal-fill design rule to meet a uniformity criterion guarantees that all blank areas greater than 1mm x 1mm in size will have the
minimal allowable density (50% in this example), it does not absolutely guarantee that the entire layout will have this value of minimum pattern density. Consider this artificial example: a layout consisting entirely of 1μm lines and 10μm spaces. Since there is no blank area greater than 50μm (which is two times the buffer distance in the above example), no regions of this chip qualify for metal-fill; but, the minimum pattern density on chip is only 10%. Also, if one corner of a layout has a large high density block and the other corner has a large low density block, the best possible range in dielectric thickness variation is ultimately dictated by the difference in the densities of these two regions. These issues are especially acute in test structures where high density defect structures (typically snakes and combs) are placed in one area and lower density circuit blocks or devices arrays are placed in another corner resulting in a severely compromised layout in the sense of only allowing marginal improvement in uniformity through metal-fill patterning. However, this situation is quite artificial and most regions of modern ASIC and logic design layout are filled with dense circuit regions requiring metal-fill only for the larger open areas. Significant improvements in uniformity are thus feasible for these designs as the case study in Section 7.5 illustrates.

### 7.4 Impact of Metal-Fill on Interconnect Capacitance

In addition to meeting a uniformity constraint, a well designed metal-fill patterning practice should also minimize the added interconnect capacitance associated with the metal-fill. The issues associated with capacitance and metal-fill are complex. Key issues are the best choice of the buffer distance (the distance between metal-fill and the nearest active metal-line), grounded\(^1\) versus floating metal-fill lines, and the shape of the metal-fill

---

1. By grounded, we mean connected to either power supply or other fixed voltage source.
patterns (e.g. lines versus blocks). For dense logic designs where the amount of metal-fill needed is low and where the number of products on which a designer is focused is also low, the placement and choice of metal-fill might ultimately be left as a decision by the circuit designer. For ASIC design, however, there are far too many designs with rapid design schedules and considerable areas which need metal-fill: automated metal-fill placement is often essential. In this section, we will discuss the trade-offs between grounded and floating metal-fill and present a methodology for developing a metal-fill design rule suitable for automated metal-fill placement with an interconnect capacitance constraint.

7.4.1 Grounded Versus Floating Metal-Fill

One of the most important decisions regarding metal-fill concerns floating metal-fill, i.e. leaving all metal-fill regions unconnected, versus grounded metal-fill, or connecting all metal-fill regions to a fixed voltage source. In terms of interconnect capacitance, grounded metal-fill tends to affect delay attributes in a layout while floating metal-fill tends to increase coupling/cross-talk attributes.

For grounded metal-fill, the primary advantage is that all metal-fill regions are at a known potential; thus, traditional layout-parasitic extraction tools can be used to re-verify and simulate a layout after the metal-fill has been placed. The drawbacks are that each metal-fill region needs to be connected to a fixed voltage source, preferably to a close-by terminal. This is often not easy and places an additional strain on already overburdened design tools. For this reason, it is better to use long lines as the metal-fill pattern and to place small metal bridges between these lines to allow ease of ground connection (see Figure 7.4a).
For floating metal-fill, the advantage is that no connections need to be made to ground; thus, floating metal-fill can be generated automatically during tape-out. The primary drawback is that floating metal-fill regions can now serve as additional coupling paths. To minimize extended range coupling effects, small unconnected square blocks should be used. Figure 7.4b illustrates this point. If vertical lines were used (as in Figure 7.5a), regions A and B, although many microns apart, could couple to each other if the lines are not

**Figure 7.4:** An illustration of vertical line filling (a) which is recommended for grounded metal-fill and block filling (b) which is recommended for floating metal-fill. Note the bridging bars that have been added in (a) to permit a connection to ground. Because the fill in (a) is grounded the metal-fill lines do not permit coupling between regions A and B although there may be an overall increase in capacitance and delay.
grounded. As Figure 7.4b shows, however, small square unconnected blocks minimize this behavior. In general, however, floating metal fill causes simulation complexities because of the uncertainties associated with floating lines. Once floating metal-fill has been introduced it is very difficult to simulate the added coupling due to floating lines in tradition parasitic extraction and verification tools as they do not (1) know how to handle floating regions separately from active regions and (2) cannot consider activity factors as part of coupling capacitance. Worst-case or (2x) coupling rules can be used, but they are overly conservative.

7.4.2 Minimizing the Effect of Metal-Fill on Interconnect Capacitance

In order to minimize the interconnect capacitance added resulting from metal-fill, the total amount of metal-fill to be added should be small, the linewidth of the fill pattern should be as small as possible, the spacing between fill “lines” should be maximum, and the buffer distance should be kept as large as possible. Unfortunately while this capacitance minimization criterion is useful as a guideline, it has two primary flaws: (1) steps restricting the amount of metal-fill and increasing the buffer distance have the unwanted effect of limiting the possible improvements in uniformity using metal-fill and (2) this minimization is not a precise method and lacks quantitative measures.

A more appropriate criterion can be formed by considering the canonical case of two lines spaced apart by twice the buffer distance plus the linewidth or block width of one metal-fill block or line. This spacing arrangement is the worst case scenario since spacing the active lines any further apart lessens any capacitance coupling/delay effects between each active lines and between any active line and any neighboring metal-fill regions. The buffer distance is initially chosen at some reasonable value (e.g. 25μm). The space
between the two lines is then patterned with metal-fill of a particular linewidth \((w)\) and linespace \((s)\) for the case of line filling (e.g. Figure 7.4a) or a particular block width \((w)\) and block space \((s)\) for the case of block filling (e.g. Figure 7.4b). A generic capacitance metric is then computed which is suitable for the particular type of metal-fill (grounded or floating).

For the floating design strategy shown in Figure 7.5a, a reasonable capacitance metric (all capacitances are per unit length) might be:

\[
C_* = \frac{\frac{C_A}{2} w}{(s + w) C_B}
\]

(7.1)

where \(C_A\) is the capacitance from an active line to a metal-fill block and \(C_B\) is the capacitance from one active line to the other active line for the case of floating metal-fill (see Figure 7.5a). For the grounded strategy shown in Figure 7.5b, a reasonable capacitance metric might be:

\[
C_* = \frac{C_A + 2C_B}{C_C + 2C_B}
\]

(7.2)

where \(C_A\) is the capacitance from an active line to a grounded metal-fill line, \(C_B\) is the overlap capacitance between an active line above and an underlying metal-fill region, and \(C_C\) is the capacitance from one active line to the other active line (see Figure 7.5b). Note that both metrics use the ratio of the capacitance (either coupling capacitance or delay capacitance) present in the metal-fill pattern to the capacitance present before metal-fill. The individual capacitance values (e.g. \(C_A, C_B, C_C\)) can either be computed using TCAD simulations or using closed-form approximations (see Appendix C). The value of \(C_*\) is
recorded and the ratio is computed again across a wide range of metal-fill spacing and width \((s, w)\) parameters.

**Figure 7.5:** A stylized illustration depicting the definitions and capacitance typically of interest in (a) floating metal-fill with block patterning and (b) grounded metal-fill with vertical line patterning.
The contour plot of $C_*$ (or $100(C_* - 1)$ which represents the percent change in capacitance) is then superimposed on top of the minimum pattern density contour plots versus line width and line space. The desired minimum pattern density specification is then coupled with the desired capacitance metric value specification to yield an optimized metal-fill design rule (see Figure 7.6). The desired capacitance metric value is selected to be as low or negative as possible, while still satisfying the desired uniformity criterion. If the capacitance metric is sufficiently small or negative, a smaller buffer length value can be selected and a comparison plot can be generated for the new buffer length value. A smaller buffer length is desirable from a uniformity perspective since smaller buffer lengths lead to denser fill and hence a higher probability of meeting the minimum pattern density specification across the entire chip. An example of the entire metal-fill methodology from uniformity criterion specification to capacitance evaluation and design rule specification is offered in the next section.
7.5 A Metal-Fill Patterning Practice Case Study

In this section, the methodology outlined in Section 7.3 and 7.4 will be exercised on the layout (see Figure 7.7) of a 256 x 32 bit 24-port memory register containing over 65,000 transistors [89]. Since completely automatic generation of metal-fill is desired, a floating metal-fill design-rule is desired. First, for the canonical structure the patterning scheme shown in Figure 7.4b will be used. The goal will be to find design rules for the three canonical parameters: the buffer distance (buf), the block width (w), and the block
space (s). Then, optimized metal-fill parameters will be selected and the impact of this optimized metal-fill on the case study (Figure 7.7) will be evaluated.

In order to estimate the effect of the metal-fill pattern on interconnect capacitance, a metric for the capacitance increase due to metal-fill as discussed in Section 7.4 needs to be determined. For this case study, we will use the canonical structure discussed in Section 7.4 and equation (7.1) to evaluate $C_\ast$. Note that for this canonical structure, the metal-fill pattern and surrounding active lines are isolated, i.e. no metal lines or features are shown above or below the structure and no ground plane has been assumed. Although in reality features are seldom isolated from each other, this canonical form has been assumed.

Figure 7.7: The metal 1 layer of the layout described in [89] used in the metal-fill case study discussed in Section 7.5. The layout is about 7.9mm by 9.2mm and the minimum linewidth and space at metal 1 is 3μm.
because it maximizes the coupling capacitance between lines (e.g. \( C_A \) and \( C_B \)). If metal features were placed above or below, some shielding would occur and the coupling capacitance between lines would be reduced by fringe capacitance from the line to the layer above or below. A solution to the more general problem is not practical since in every area of a layout the metal coverage above and below a metal-layer varies significantly. A statistical approach could be attempted in which the average and range of patterns above and below a given layer are extracted and used to form a model database for capacitance evaluation, but this technique would complicate the formulation of a design rule since the capacitance metric (c.f. \( C_* \) above) becomes a distribution rather than one number.

The large number of calculations required (> 1000) to compute the individual components of \( C_* \) (e.g. \( C_A \) and \( C_B \)) and generate design-rule charts motivates the use of simple closed form expressions instead of numerical simulations. More specifically, we will assume that \( C_A \) and \( C_B \) (see Appendix C) can be approximated by:

\[
C_B \propto \left( \frac{t}{buf} \right)^n
\]

(7.3)

\[
C_A \propto \left( \frac{t}{buf} \right)^n + 2t^n \ln \left[ \frac{bs}{2} + \frac{bs^2}{2} + \left( buf + bw \right)^2}{buf + \frac{bw}{2}} \right]
\]

(7.4)

where \( t \) is the thickness of the metal and \( bs \) is the spacing between blocks, \( bw \) is the block width, and \( buf \) is the buffer length (see Figure 7.5a). As Figures 7.8a and 7.8b show, the approximations in (7.3) and the assumption of linewidth independence is quite good except for small spaces where (7.3) overestimates the capacitance. Also note that the proportionality constant in (7.3) is primarily dependent on metal thickness and not signifi-
cantly affected by linespace. For the comparison in Figure 7.8, the two-dimensional capacitance solver RAPHAEL™ [90] was used to evaluate the capacitance; the simulation structure is shown in Figure 7.8c.

Figure 7.8: a) Interconnect capacitance per unit length for two parallel lines 0.6 micron thick. Note the relative independence of capacitance to linewidth. (b) The effect of interconnect capacitance for two parallel lines versus spacing and metal thickness with the metal width fixed at 2.5 μm. Note the relatively good agreement ($R^2 > 0.99$) between the simulated capacitance values and the model shown and discussed in Section 7.5. (c) The simulation structure used in the comparison shown in (a,b).
In (7.4), the capacitance $C_A$ is composed of two terms. The first term accounts for the coupling between the sidewalls of the active line on the left (line $b_1$ in Figure 7.9) and one of the metal-fill blocks. The second term represents fringing from the sidewall of line $b_1$ to the perpendicular sidewall of a metal-fill block. A derivation for this second fringing term can be found in Appendix C. In order to assess the validity of the assumptions and approximations in (7.4), several three-dimensional simulations were run using RAPHAEL™ [90] for the structure shown in Figure 7.9. As Table 7.2 shows, the agreement between

**Figure 7.9:** The simulation structure used to evaluate the capacitance approximation in (7.4). For the comparison, all lines except $b_1$ were grounded, and the capacitance between $b_1$ and every other metal region was simulated using a three dimensional capacitance solver.
simulation and (7.4) is excellent. For the comparison shown in Table 7.2, all lines except \( b_1 \) were grounded, and the capacitance between \( b_1 \) and every other metal region was simulated.

**Table 7.2: Comparison of the capacitance approximation in (7.4) versus simulation.**

<table>
<thead>
<tr>
<th>buffer length (µm)</th>
<th>block width (µm)</th>
<th>block space (µm)</th>
<th>Capacitance (sim.) (fF)</th>
<th>Capacitance (fF) [from (7.4)]</th>
<th>% Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0.099</td>
<td>0.096</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>1</td>
<td>0.041</td>
<td>0.041</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0.185</td>
<td>0.182</td>
<td>1.6</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0.137</td>
<td>0.148</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>0.180</td>
<td>0.188</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
<td>0.147</td>
<td>0.138</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 7.10 shows a contour plot of \( C_v \) versus block width and block space superimposed on a contour plot of the minimum pattern density within a 1mm block (see Section 7.3) versus block width and space. Contour plots are shown for buffer lengths of 10µm, 25µm, 50µm and 100µm. If an entire layout were composed of features at the minimum feature width and space for metal-1, then the pattern density across the chip should average around 50%. Thus for our case study, we select the minimum pattern density criterion to be 50% since the majority of features in the layout in Figure 7.7 are placed at minimum width and space.
Table 7.3 shows the block width and space combinations which achieve the 50% minimum pattern density criterion. These values were extracted from Figure 7.10 using a minimum feature width and space of 3μm as dictated by process considerations at the time Figure 7.7 was designed. For Table 7.3, larger choices of line space invariably lead to higher capacitance metric numbers; thus, the minimum values were chosen (see Figure 7.10: Design charts of minimum pattern density constraints superimposed on a relevant metric for evaluating the effect of a particular metal-fill design rule on interconnect capacitance. Design charts are shown for buffer lengths of 10μm, 25μm, 50μm, and 100μm. These charts allow one to contrast uniformity constraints with capacitance requirements. In reality, many of the negative capacitance numbers shown are slightly larger due to neglected fringing elements.)
Although the capacitance metric minimizes for a buffer length closer to 50μm, a buffer length of 25μm was chosen, because a 50μm buffer length would limit the minimum space that could be filled by metal-fill to over 100μm (which would limit the permissible change in pattern density) and the relative gain in the capacitance metric numbers would be marginally small. Note that if a different minimum feature width and space is assumed, the numbers in Table 7.3 change, but the optimal buffer length is still in the 25 to 50μm range.

**Buffer Length = 25μm**

![Contour Plot of Pattern Density and Capacitance Metric](image)

*Figure 7.11: Detailed view of the metal-fill design chart for a buffer length of 25μm. For the case study discussed in Section 7.5, the optimal choice of 9μm and 3μm for linewidth and linespace is highlighted. This value achieves a minimum pattern density of 50% and minimizes any added capacitance associated with the metal-fill. Linewidths and spaces below 3μm were not considered manufacturable.*
Figure 7.12 shows the simulated dielectric thickness between metal-1 and metal-2 using the model in [59] for the layout shown in Figure 7.7. Figure 7.12 (a,b) shows the ILD thickness variation before metal-fill patterning and Figure 7.12 (c,d) shows the ILD thickness variation after optimized metal-fill patterning using a buffer length of 25\(\mu\)m, a block width of 9\(\mu\)m, and a block space of 3\(\mu\)m. The range in ILD thickness variation has been reduced from 0.22\(\mu\)m to just under 0.04\(\mu\)m, a reduction of 82%. In addition to removing almost completely the impact of dielectric thickness variation on circuit performance, substantial gains have been made towards the manufacturability of this product (e.g. dielectric thickness variation has been factored out of depth-of-focus error budgets.)

<table>
<thead>
<tr>
<th>Buffer Length ((\mu)m)</th>
<th>Block Width ((\mu)m)</th>
<th>Block Space ((\mu)m)</th>
<th>% Capacitance Increase</th>
<th>Minimum Space that can be Filled by Metal-Fill Regions ((\mu)m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8</td>
<td>3</td>
<td>13%</td>
<td>31</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>3</td>
<td>-1%</td>
<td>62</td>
</tr>
<tr>
<td>50</td>
<td>11</td>
<td>3</td>
<td>-5%</td>
<td>114</td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>3</td>
<td>0%</td>
<td>228</td>
</tr>
</tbody>
</table>
Conclusion

In this chapter, a methodology has been demonstrated for developing design rules for metal-fill patterning practices which take into account dielectric thickness uniformity constraints as well as the effect of metal-fill patterning on interconnect capacitance. The methodology presented in this chapter represents an important tool for reducing dielectric thickness variation in CMP processes. The procedure outlined here is of particular interest to the ASIC community where the number of products manufactured and the variety of
layouts and designs encountered are both large necessitating an automated procedure for mitigating dielectric thickness variation.

Several extensions of this work can be identified. Most notably, the methodology presented in this paper can be extended to shallow trench isolation processes and inlaid metal techniques such as copper damascene. Also, the methodology might also be adapted to spin-on-glass (SOG) or to other novel dielectrics processes. Finally, novel metal-fill patterning procedures can be developed. Key novel procedures include (1) attaching all metal-fill patterns to ground gated through transistors so that the amount of grounding or floating can be modulated as a function of switching activity or circuit cell function and (2) activity dependent metal-fill patterning in which metal-fill is used aggressively near non-critical slow transitioning circuit blocks while conservative patterning is used near critical paths and rapid switching circuit blocks.
Chapter 8

Conclusion

8.1 Statistical Metrology as Design for Manufacturability

In many ways, the methodology presented in this work represents an application of design for manufacturability (DFM). The interplay between design and manufacturing is perhaps most striking in variation issues. The analysis procedures can also serve as yield/performance drivers. Clearly, wafer-scale variation effects such as poly-CD variation impact the final speed and yield of a product in manufacture. Understanding the root causes of this variation drives methods for optimizing the process towards less variability and hence more performance or yield. Also, devices and interconnect process are often designed around manufacturing constraints. Often, the fastest devices and interconnect processes are the most unmanufacturable. A reduction in variation or an understanding of its character enables many of these architectures or processes more viable. Finally, the methodology presented in this work moves the design process closer to first pass prediction. Currently, several design iterations are necessary on newly introduced products because these products fail to function or glitch too frequently to meet desired performance goals as a result of an inability to predict inevitable fluctuations of variation signatures in the “fab” Learning more about the underlying physics at work in variation and how this variation impacts performance allows expensive design/manufacturing interactions to be traded for far less expensive tactics such as simulation and data analysis.

In this thesis, a novel methodology has been developed for characterizing variation in semiconductor manufacturing with special emphasis on pattern dependent and spatial variation. Specific methods for identifying the nature and scope of variation, modeling the variation as a function of layout and/or process parameters, and determining the impact of
pattern dependent variation on circuit performance are contributed. Illustrations of this methodology were drawn from key interconnect and device variation concerns with particular concern for ILD thickness variation in oxide CMP processes. In this chapter, we identify area for future work and briefly encapsulate the contributions of each chapter.

8.2 Variation Assessment
In the variation assessment phase of this work, two methods for decomposition spatial variation into wafer-level, die-level, interaction-terms, and residuals has been demonstrated. Several extensions and recommendations for future work can be identified.

- The meshed spline method for extraction wafer-level variation is still not fully understood. When exactly does the meshed-spline method outperform the down-sampled moving average? Instead of a pseudo-2D spline smooth, can true 2-D splines such as the thin-plate spline be used?
- The method used to select the downsampling rate in the DSMA extractor is still unrefined and could use a more analytical approach rather than a data driven procedure.
- Is the regression based wafer-level estimator always the worst estimator or does it sometimes outperform even the DSMA estimator?
- What is the role of missing die or data on the accuracy of the FFT based die-level estimator?
- For the interaction terms estimator, what should be done for very non-spatially stationary signals?
- Can a 2-D spline improve the performance of the spline based interaction term estimator?
- Can the methodology developed in Chapter 2 be extended for spatially based lot-to-lot and wafer-to-wafer variation such as for thin films grown in a furnace? An example of this type of variation was presented in [15, 16].
- Can we generalize when the spline based interaction term estimator is better than the FFT based method and vice-versa?
- What is the effect of stepper leveling errors (i.e. variation which is systematic within a die but random from die-to-die) and Poisson type variation on the estimators presented in Chapter 2? Can these types of variation signatures be extracted?
• A better method for rapidly extracting the wafer-level variation other than the single structure method is needed.

• The multiplicative model interaction term estimator is still not completely understood. When will this model fail? Can we devise a statistical test to determine the efficacy of a multiplicative model?

• Finally, how do temporal disturbances factor into decomposition analysis. Can temporal effects be blocked by averaging all the wafers together from one lot to form one “average” wafer?

8.3 Variation Modeling
Two approaches to variation modeling were presented in Chapters 4 and 5. The methodology espoused in Chapter 4 explored investigative modeling for cases when little or no expectations about the size or nature of the pattern dependencies exist. Several unresolved and unexplored issues are:

• Better methods are need for integrating extraction of wafer-level variation (which are often non-parametric estimators) with the die-level models (which are often parametric). Clearly, the regression based approach to estimate the wafer-level surface is non-optimal.

• Using MANOVA based models (which assuming underlying normal distributions) with fundamentally non-normal and systematic variation is probably not correct and this issue needs to be revisited.

• A more thorough investigation into the physical cause of many poly-CD and metal-CD issues is warranted.

• Can the techniques presented in Chapter 5 be applied to other variation types such as metal CMP or other types of parametric variation?

8.4 Variation Impact
In Chapter 6, a methodology for determining the impact of pattern dependent variation on circuit performance was illustrated. This methodology should enable a more complete determination of just what is the impact of pattern dependent variation on circuit performance. Other related key issues are:

• The effect of scaling and circuit design styles. Do all circuit architectures for adders, for example, respond similarly to pattern dependent variation or is one design style more preferred. Is one clocking style more robust to pattern dependent variation then another style?
• In Chapter 6, the “rule-of-thumb” that the degree to which variation is correlated along a net appears to dictate the sensitivity of that net is purely anecdotal and based on the case studies. Clearly, this rule needs further verification.

• What about the role of pattern dependent variation on analog circuits? Can any results be generalized?

• Finally, the use of this methodology for determining impact still needs to be validated with a real product or full-flow silicon.

### 8.5 Variation Reduction

In Chapter 7, the use of metal-fill patterning for variation reduction was presented. Several extensions which can be identified are:

• The use of metal-fill for variation reduction in STI process and for use in metal damascene technologies.

• A closer examination of the role of metal-fill on capacitance.

Finally, the entire concept of variation reduction is open for research. Clearly, techniques such as optical proximity correction (OPC) are suspect topics; however, other nascent variation reduction techniques need to be identified.
Appendix A

Artificial Dataset

Figure 2.4 shows the artificial dataset used to evaluate the relative and absolute performance of the wafer-level (Figure 2.4b), die-level (Figure 2.4c), and wafer-die interaction (Figure 2.4d) estimators. The wafer-level variation model shown in Figure 2.4b was generated using the formula:

\[
L_V = 10 - \left\{ \frac{a(x - 30)^2 + b(y - 30)^2 + c x y + d x + e y + f \cos(x + y)}{300} \right\}
\]

where \(a, b, c, d, e,\) and \(f\) are adjusted to create the wafer-level variation. For the wafer-level model shown in Figure 2.4b, the coefficients in (A.1) were chosen arbitrarily at \(a = 1, b = 1, c = 0.2, d = 0.4, e = 3,\) and \(f = 0.2\) and \(x\) and \(y\) were defined to be the integers 1 through 60. For the die level variation, the model shown in Figure 2.4c was empirically generated based on data on intra-die polysilicon critical dimension variation found in [15]. The values for one die in the die-level variation can be found in Table A.1. The wafer-die interaction terms shown in Figure 2.4d were generated using the transformation:

\[
f_{WLV @ DL V} = 0.1 \times f_{DLV} \times f_{WLV} \times l_1 \times l_2
\]

where

\[
l_1 = \begin{cases} 
1 - \left| \frac{x - 30}{30} \right| & |x - 30| > 15 \\
0.5 & otherwise
\end{cases}
\]

and

\[
l_2 = \begin{cases} 
1 - \left| \frac{y - 30}{30} \right| & |y - 30| > 15 \\
0.5 & otherwise
\end{cases}
\]

Finally, random zero-mean additive Gaussian noise with a standard deviation of 0.01 was added to form the model shown in Figure 2.4a.
### Table A.1: Die-Level Variation Data (One Die Shown)

<table>
<thead>
<tr>
<th></th>
<th>1.1</th>
<th>1.4</th>
<th>1.5</th>
<th>1.4</th>
<th>1.1</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>1.4</td>
<td>1.5</td>
<td>1.4</td>
<td>1.1</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>1.7</td>
<td>1.8</td>
<td>1.7</td>
<td>1.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.8</td>
<td>1.9</td>
<td>1.8</td>
<td>1.5</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>1.7</td>
<td>1.8</td>
<td>1.7</td>
<td>1.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>1.4</td>
<td>1.5</td>
<td>1.4</td>
<td>1.1</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.9</td>
<td>1.0</td>
<td>0.9</td>
<td>0.6</td>
<td>0.1</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

Spline Smoothing Theory

Many statistically based experiments use regression in one form or another. Generally speaking, there are two main types of regression analysis -- parametric and nonparametric regression. Parametric regression assumes that a model form is known or given for the output of interest except for a finite number of unknown parameters. Linear least squares estimators and polynomial regression are common examples of parametric regression. Nonparametric regression is a more general technique. Nonparametric regression models assume that the output of interest belongs to some infinite dimensional collection of functions such as the set of all twice differentiable functions. Spline smoothing is one particular method used to implement nonparametric regression.

There are four main types of spline smoothers: polynomial regression splines, natural splines, cubic splines, and cubic B splines. Polynomial regression splines fit a polynomial of a specified degree piecewise to the data. The piecewise regions are separated by breakpoints, or knots, and the individual pieces must fit together smoothly at these knots. Natural splines are polynomial splines of any odd degree with the added constraint that the function must be linear beyond the boundary knots. Cubic splines are piecewise cubic polynomials which are continuous and twice differentiable at the knots. Cubic B splines are a more generalized and robust form of cubic splines, and are summarized below. The spline smoother used to extract the wafer-die interaction terms and in the multiple spline method wafer-level estimator are cubic B splines.

Cubic splines operate by seeking the twice-differentiable function, $f(x)$, which minimizes the penalized residual sum of squares (PRSS)
The first term is a sum of squares error which penalizes deviations from the data while the second term penalizes curvature in the estimator function. The smoothing parameter $(SPAR)$, or $(\lambda)$, controls the shape of the weight function. Large values of the smoothing parameter lead to smoother curves. As $\lambda \to \infty$ the second term dominates and the resulting estimator is the same solution extracted from a linear least squares estimator. As $\lambda \to 0$, the solution tends to a twice-differentiable interpolating function [18].

In a cubic spline smoother, the basis of the estimator function is a truncated power series basis

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{B.2}$$

In contrast, the cubic B spline smoother uses the B spline basis functions which are defined recursively. Assuming there are an infinite number of knots, $... < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < ...$, the B spline of degree 0 is defined as:

$$B_i^0 = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{else} \end{cases} \tag{B.3}$$

and B splines of higher degree are defined as:

$$B_i^k(x) = \left( \frac{x-t_i}{t_{i+k}-t_i} \right) B_i^{k-1}(x) + \left( \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} \right) B_{i+1}^{k-1}(x). \tag{B.4}$$

It can be shown that these functions form a better basis compared to (B.2) as well as have several other properties [19, 34].

Although the smoothing parameter is often difficult to estimate, three common techniques yield reasonably good estimates. The first and most unreliable method is trial and error in which a value for $\lambda$ is chosen and the resulting smooth is observed. If the curve is
not sufficiently smooth, the process is repeated. The second technique is by cross-validation. The mathematics and theory behind this method are beyond the scope of this work[21, 24], but it generally achieves a reasonably good estimate for the smoothing parameter. The final technique involves choosing the degrees of freedom. The degrees of freedom is related to the number of knots [24]. In this way, the relative smoothness or roughness of a fit can be controlled to in essence elicit variation associated with different ranges of interaction and hence physical phenomena. Other smoothers and nonparametric regression models exist but the cubic B spline smoother is preferred in the algorithms presented in this paper because an association between smoothness/roughness, degrees of freedom, ranges of interaction, and physical processes can be established.
Appendix C

Approximations for the Capacitance from Metal-Fill Regions

Consider the illustrative drawing shown in Figure C.1. If we assume a parallel plate capacitance formula similar to (7.3) holds then the capacitance (per unit length) for the case shown in Figure C.1 can be written as:

\[ dC_1 = a \left( \frac{t}{\sqrt{x^2 + (buf + bw/2)^2}} \right)^n dx = a \frac{t^n}{\sqrt{x^2 + (buf + bw/2)^2}} dx \quad \text{for } n = 1 \quad (C.1) \]

where \( a \) is a proportionality constant. In (C.1) we are also assuming that \( buf >> bw/2 \). For the examples considered in this paper, this is a valid assumption. The capacitance per until length in (C.1) can then be integrated across the entire span of \( x \) from 0 to \( bs/2 \):

\[
C_1 = a \int_0^{bs/2} \left( \frac{t^n}{\sqrt{x^2 + (buf + bw/2)^2}} \right) dx \quad (C.2)
\]

Integrating (C.2), we obtain:

\[
C_1 = at^n \ln \left[ \frac{bs}{2} + \sqrt{\left( \frac{bs^2}{2} + (buf + bw/2)^2 \right)} \right] \quad (C.3)
\]

for the fringing component, \( C_1 \), shown in Figure C.1. For the coplanar capacitance, \( C_2 \), the formula shown in (7.3) can be used. Summing \( 2*C_1 \), to account for fringing from both edges, and \( C_2 \) gives the approximation of (7.4).
Figure C.1: An illustrative diagram used in the calculation of the fringing capacitance formula discussed in Section 7.5.
References.


dexter, C. Pollard, G. Ross, J. Ryan, S. Wolff, J. Cronin, VLSI Multilevel Intercon-  


Mechanical Polishing (CMP),” VLSI Multilevel Interconnect Conference, pp. 464-  
470, Santa Clara, CA, June 1995.

Processes,” VLSI Multilevel Interconnect Conference, pp. 57-63, Santa Clara, CA,  
June 1990.

Planarization Rates Using Atomic Force Microscopy,” Chemical Mechanical Pol-  
ishing of VLSI/ULSI Multilevel Interconnect Conference, pp. 74-81, Santa Clara, CA,  
June 1996.

Characterization,” VLSI Multilevel Interconnect Conference, pp. 438-441, Santa  
Clara, CA, June 1990.

Kapoor, “The Role of Dummy Fill Patterning Practice on Intra-Die ILD Thickness  
Variation in CMP Processes,” VLSI Multilevel Interconnect Conference, pp. 421-  
423, Santa Clara, CA, June 1996.

[53] B. Stine, D. Boning, J. Chung, “Analysis and Decomposition of Spatial Variation in  
1, February, 1997.

Variation,” Microelectronic Manufacturing Yield, Reliability, and Failure Analysis  
TX.


Effects in the Chemical-Mechanical Polishing of Inlaid Copper Structures,” J. Electro-  


Oh, S. Prasad, W. Loh, A. Kapoor, “Assessing and Characterizing Inter- and Intra-  
die Variation Using a Statistical Metrology Framework: A CMP Case Study,” First  
Intl. Workshop on Statistical Metrology, Honolulu, HI, June 1996.


