Modeling the Effects of Systematic Process Variation on Circuit Performance

by

Vikas Mehrotra

Bachelor of Science, Wright State University, 1993
Master of Science, The Ohio State University, 1995

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

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 11 2001
LIBRARIES
BARKER

May 2001
[June 2001]

© 2001 Massachusetts Institute of Technology. All Rights Reserved.

Author ....
Department of Electrical Engineering and Computer Science
May 21, 2001

Certified by ....
Duane S. Boning
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

Modeling the Effects of Systematic Process Variation on Circuit Performance

by

Vikas Mehrotra

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2001 in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Electrical Engineering and Computer Science

Abstract

As technology scales, understanding semiconductor manufacturing variation becomes essential to effectively design high performance circuits. Knowledge of process variation is important to optimize critical path delay, minimize clock skew, and reduce crosstalk noise. Conventional circuit techniques typically represent the interconnect and device parameter variations as random variables. However, recent studies have shown that strong spatial pattern dependencies exist, especially when considering interconnect variation in chemical mechanical polishing (CMP) processes. Therefore, the total variation can be separated into systematic and random components, where a significant portion of the variation can be modeled based on layout characteristics. Modeling the systematic components of different variation sources and implementing these effects in circuit simulation are key to reduce design uncertainty and maximize circuit performance.

This thesis presents a methodology to incorporate systematic pattern dependent interconnect and device variation models for use with circuit extraction and simulation tools. The methodology is applicable to variation impact assessment as well as variation reduction during circuit design. Systematic models are implemented within a computer aided design (CAD) tool environment to enable automated analysis since the impact of variation is a function of circuit type, performance metric, type of technology, and type of variation source under consideration. The methodology is then applied to study the effects of different variation sources on high performance microprocessor circuit designs for the various performance metrics. The impact of variation is also projected as technology is scaled to the 50 nm generation. Our results indicate that design margin can be tightened significantly if systematic variation models are used for circuit simulation, especially with technology scaling.

Thesis Supervisor: Duane S. Boning

Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my thesis supervisor, Professor Duane Boning, for his guidance on this research. I am grateful for his advice throughout the course of this project and the valuable suggestions on this thesis. I would also like to thank the thesis readers, Professor Anantha Chandrakasan, Professor Jacob White, and Dr. Sani Nassif of IBM for serving on my thesis committee. In addition, I want to thank Dr. James Chung, now with Compaq, who provided many of the ideas early on in this work.

This thesis is the result of significant industrial collaboration. I would like to thank all the people I worked with at IBM Austin Research Lab, Austin, TX, and PDF Solutions, San Jose, CA. IBM provided many of the circuit designs for the case studies in this thesis as well as access to their circuit design and simulation tools. I also thank PDF Solutions for access to their CAD tools to help develop the methodology for this thesis. I especially want to thank Sani Nassif at IBM and Rakesh Vallishayee at PDF Solutions. I also thank Sam Nakagawa of Hewlett Packard for providing the data for one of the case studies.

I want to thank all my colleagues from the Statistical Metrology group and MTL for the interesting conversations and stimulating research discussions. Specifically, I would like to the thank the following people, including both current group members and recent graduates: Tae Park, Tamba Tugbawa, Brian Lee, Dennis Ouma, Brian Stine, Rajesh Divecha, Huy Le, Arifur Rahman, Taber Smith, Jung Yoon, Jee-Hoon Krska, Shiou Lin Sam, Aaron Gower-Hall, Karen Gonzalez-Valentin, Kuang Han Chen, Joseph Panganiban, David White, Allan Lum, and Mike Mills.

I also want to thank my family for their encouragement and support over the past several years during my Ph.D. research at MIT.

This work has been supported in part by grants from DARPA, PDF Solutions, and the MARCO Interconnect Focus Center.

Table of Contents

1 **Introduction** 21

1.1 Sources of Variation 21

1.1.1 Interconnect Variation 22

1.1.2 Device Variation 22

1.2 Motivation 23

1.2.1 Systematic vs. Random Variation 26

1.2.2 Systematic Variation in Interconnect CMP Processes 30

1.2.3 Interconnect Variation Impact on Circuit Performance 34

1.2.4 Technology Scaling Impact 36

1.3 Thesis Goals 38

1.4 Thesis Organization 40

2 **Circuit Performance Metrics** 41

2.1 Interconnect Modeling 41

2.1.1 Capacitance Calculation 42

2.1.2 Resistance Calculation 45

2.2 Signal Delay 45

2.3 Clock Skew 50

2.4 Crosstalk Noise 51

2.5 Inductance Modeling 53

2.5.1 Inductance Calculation 55

3 **Process Variation Models** 57

3.1 An ILD CMP Model 58

3.2 A Copper CMP Model 61

3.3 Critical Dimension Variation 63

3.4 Example: Systematic vs. Worst-Case Variation Modeling.................................66

4 Systematic Variation Analysis Methodology.......................................71

4.1 Variation Analysis Simulation Methods..71

4.2 Net Extraction 73

4.3 New Methodology for Variation Assessment..76

4.3.1 Pattern Density Calculation 81

5 Systematic Variation Impact Assessment..85

5.1 ILD Thickness Variation Impact on 1 GHz Microprocessor..............................85

5.1.1 Computing ILD Thickness Variation 86

5.1.2 Effect of Metal Fill 87

5.1.3 Interconnect Capacitance Distribution 89

5.1.4 Interconnect Delay Distribution 89

5.2 Copper CMP and Poly CD Variation Impact on 1 GHz Clock Tree..................93

5.2.1 Interconnect Variation Modeling 93

5.2.2 Poly CD Variation Modeling..93

5.2.3 Clock Driver and Latches 95

5.2.4 Clock Skew 96

5.3 Interconnect Delay Variation: Metal vs. Oxide CMP...98

5.3.1 Copper CMP Interconnect Variation...100

5.3.2 Oxide CMP ILD Thickness Variation..100

5.3.3 Interconnect Delay Variation 101

5.3.4 Discussion of Interconnect Variation Impact on Delay 103

5.4 Summary 104

6 **Systematic and Random Variation Effects** **105**

6.1 Clock Skew in H-Tree 106

6.1.1 Metal Thickness Variation 106

6.1.2 Device and Power Supply Variation 106

6.1.3 Clock Skew 108

6.2 Optimally Buffered Interconnect 110

6.2.1 Buffer Insertion and Optimal Interconnect Delay 110

6.2.2 Interconnect Length Variation 112

6.2.3 Delay Variation 115

6.3 Impact of Variation on Crosstalk Noise 118

6.4 Variation Impact on Crosstalk With Optimal Interconnect 122

6.5 Summary 124

7 **Technology Scaling Impact** **127**

7.1 Technology Scaling 127

7.2 Technology Scaling Impact on Clock Skew 128

7.2.1 Clock Skew 130

7.3 Interconnect and Device Variation Scaling 131

7.3.1 Interconnect Variation Scaling Scenarios 134

7.3.2 Device Variation Scaling 136

7.4 Variation Impact With Technology Scaling 137

7.4.1 Effect of Scaling Scenarios on L_{max} 137

7.4.2 Worst-Case Models vs. Tighter Tolerance Design 138

7.4.3 Scaling Scenario Comparisons: Effect on L_{max} 141

7.5 Technology Scaling Impact on Crosstalk Noise 143

7.6 Summary 144

8 **Conclusion** **147**

8.1 Summary 147

8.2 Conclusion 148

8.3 Future Work 149

Bibliography **151**

List of Figures

Figure 1.1: Cross-section of parallel interconnect lines above a ground plane. The figure on the left (a) shows the ideal case, and the figure on the right (b) shows some of the different types of variation that can exist in the interconnect. 22

Figure 1.2: Cross-section (a) and top view (b) of an NMOS device on a silicon wafer. Structural variations in the polysilicon and gate oxide impact the performance of the device in addition to variations in the active areas or doping profiles. 23

Figure 1.3: The input variables X_1 and X_2 are randomly distributed. If the inputs are correlated, the samples may lie within the region bounded by the ellipse. In this case, using minimum and maximum limits can result in pessimistic corners. 25

Figure 1.4: The output distribution is obtained by taking random samples of the inputs X_i and performing several simulations (typically thousands). The input corners can be obtained by determining the values of the inputs such that the output is inside the specified confidence interval (i.e. find bounds for X_i such that $Y_{min}<Y<Y_{max}$). 25

Figure 1.5: Variation can be decomposed into different components, including lot-to-lot, wafer-to-wafer, within wafer, and intra-die. The intra-die component can be a large component of the total variation. For example, the most significant source of interconnect ILD thickness variation is the intra-die component 27

Figure 1.6: The variation is separated into components using a series of estimators. After each stage, a variation component is extracted and the residuals are fed to another estimator for extracting the next component [17]. 28

Figure 1.7: Random sampling within a single die is used to determine the statistics for input variable X_1 (a). Each structure within the die is sampled across multiple die within a given wafer for several wafers (b). 29

Figure 1.8: A rotary CMP tool. Both chemical and mechanical action is used in polishing the wafer. The carrier and platen rotate as a slurry chemical is fed to the polishing pad. ..32

Figure 1.9: A short flow process where the metal is patterned, followed by oxide deposition and CMP. In the ideal case, the dielectric is polished so that a smooth and flat surface results after CMP. 32

Figure 1.10: The ILD thickness variation across any individual die can be separated into die-level, wafer-level, wafer plus die, and random components. The greatest amount of non-uniformity is in the die-level component, indicating the importance of pattern dependence. 33

Figure 1.11: Post CMP ILD thickness is a strong function of underlying metal pattern density. With conformal deposition, sparse regions polish faster than dense regions, resulting in a greater amount of oxide above dense areas. 33

Figure 1.12: Interconnect delay is generally a large fraction of the total signal delay for critical paths. Variations in the interconnect impact its electrical properties and can limit circuit performance. 34

Figure 1.13: An H-tree used for clock distribution. Wires in the tree are tapered for impedance matching [2]. When a fork is encountered, the width of the next branch is reduced by a factor of two. 35

Figure 1.14: The input signal V_{in2} is quiescent, but its neighbors are switching from low to high. Due to the lateral coupling between lines, crosstalk noise causes a spike at V_2.......36

Figure 1.15: A three-tier interconnect wiring scheme typically found in high performance circuit designs. Each tier contains two or more wiring levels, including alternating levels for horizontal and vertical routing. Wires in the intermediate and global tiers are used for long distance routing. Power supply and clock lines generally use global wires. 37

Figure 1.16: The maximum distance that can be routed using minimum pitch local, intermediate, and global interconnect to meet the global clock frequency constraint for different technology generations without intermediate buffers (a) and with buffer insertion (b). By the 100 nm generation, the longest global wires will require more than one clock cycle even with intermediate buffers. 38

Figure 2.1: The pi (a), T (b), and ladder (c) representations of the interconnect. A long wire is broken down into N sections for circuit simulation to account for the distributed effect... 42

Figure 2.2: A distributed RLC representation of the interconnect network. The inductance must be included at high frequencies. 42

Figure 2.3: The interconnect capacitance is a function of wire geometry. It includes ground capacitance, lateral coupling, and fringing effects (a). The interconnect can be modeled as a series of distributed RC sections that include separate line-to-ground and line-to-line coupling capacitances (b). 44

Figure 2.4: The signal delay is modeled using the circuit shown above. The total path delay includes the effects of the driver resistance, device load capacitance, and the distributed RC interconnect load. 46

Figure 2.5: The CMOS inverter (a) is modeled as a switch with a finite on-resistance (b) for computing the signal delay. The NMOS resistance is used for a low to high input transition and the PMOS resistance is used for a high to low output transition. 48

Figure 2.6: The different components of the device load capacitances. The load seen at the driver output is the sum of the gate-to-drain and drain-to-bulk capacitances of the input buffer plus the gate capacitances at the output buffer and the distributed RC interconnect load. 49

Figure 2.7: The clock distribution network using a spine configuration (a) automatically causes skew due to a difference in the interconnect lengths between the clock driver and the outputs. The H-tree (b) has clock skew because of a difference in the loads, although the distance between the driver and all loads is the same. Both cases are examples of clock skew due to design asymmetry. 52

Figure 2.8: An aggressor signal switches next to a quiescent line (a) and results in crosstalk noise at the victim output. A 2D cross-section showing the capacitances of the victim and aggressor signals (b) and the corresponding equivalent circuit used to compute the maximum crosstalk noise $V_{2,\,max}$ (c). 54

Figure 3.1: The oxide polish rate depends on the underlying metal pattern density until the oxide features ("up areas") have been removed. In this regime, there is negligible polishing of the "down areas". In the second regime (when all up areas are removed), the oxide polishes at the blanket removal rate. 60

Figure 3.2: The local density is taken as the ratio of linewidth to pitch (a). The effective pattern density depends on the planarization length (b). A larger interaction distance results in more averaging across the chip. 60

Figure 3.3: A damascene CMP process used with Cu interconnect. The oxide is first patterned, followed by metal stack deposition. The metal is polished until all "up areas" have been removed. To guarantee that there are no inadvertent shorts between adjacent wires, overpolishing is done resulting in metal thickness loss. 62

Figure 3.4: Overpolish in copper CMP processes results in metal dishing and oxide erosion. Dishing is present in wide lines and erosion dominates for fine pitch lines at high density, both causing metal thickness loss. 63

Figure 3.5: Metal dishing (a), oxide erosion (b), and total metal thickness loss (c) as a function of linewidth and linespace. Erosion is the dominant component for the range of *LW* and *LS* combinations shown here. 64

Figure 3.6: The pattern density (a) and ILD thickness variation (b) between the top two metal layers in a microprocessor using an interaction distance of 4 mm. The simulated delays of 100 interconnect lines 10 mm in length placed randomly across the chip (c) and a histogram showing the distribution (d). The worst-case ILD thickness value overestimates the actual worst-case by 8%....................69

Figure 3.7: The pattern density (a) and ILD thickness variation (b) between the top two metal layers in an ASIC using an interaction distance of 4 mm. The simulated delays of 100 interconnect lines 10 mm in length placed randomly across the chip (c) and a histogram showing the distribution (d). The worst-case ILD thickness value overestimates the actual worst-case by 26%. 70

Figure 4.1: The conventional approach to modeling the effects of variation using statistical circuit analysis. Variation statistics are input for each parameter and Monte-Carlo simulations are performed to obtain a delay distribution. 73

Figure 4.2: In addition to the net capacitance or resistance, information about the geometry and coordinates of each net and its surrounding neighbors is also stored during the extraction. This information is then used to predict the spatial variation (e.g. ILD or metal thickness variation) and compute the change in electrical parameters. 75

Figure 4.3: The methodology used to simulate the effects of pattern dependent interconnect and device variation on circuit performance. Elements in bold are used to model the effects of variation. 79

Figure 4.4: The procedure used to implement systematic variation models in a CAD tool and automate performance impact assessment. 80

Figure 4.5: A moving average square window with planarization length *L* is used to compute the effective pattern density at different points on the die. 81

Figure 4.6: The effective pattern density for the Metal 5 layer of a microprocessor using an interaction distance of 3.5 mm (a) and 100 μm (b). 83

Figure 4.7: Simulated ILD thickness obtained by setting the polish time for a nominal target ILD thickness of 1 μm at 50% pattern density (a). The ILD thickness obtained by re-centering the polish time based on average effective pattern density (b). The range of variation remains the same, while the maximum deviation from the 1 μm target is reduced. 84

Figure 5.1: The 1 GHz microprocessor used in simulating the effects of global path delay due to ILD thickness variation. 86

Figure 5.2: Adding metal fill reduces the ILD thickness non-uniformity by increasing the effective pattern density across the chip. 88

Figure 5.3: The global net capacitance distribution without (a) and with (b) metal fill. The addition of metal fill reduces the maximum capacitance variation from 28% to 18% (compared with nominal values). 90

Figure 5.4: The global path delay variation for the microprocessor shown in Fig. 5.1. The delay variation is simulated with the pattern dependent ILD thickness variation model (a), assuming enough metal fill is added to limit the minimum pattern density at 30% (b), and the worst-case limits on ILD thickness (c). 91

Figure 5.5: Summary of the interconnect delay variation for the various cases in Fig. 5.4... 92

Figure 5.6: The clock tree used to simulate the effects of copper CMP variation on skew. The tree is driven by a series of cascaded drivers at the input and loaded with latches at the outputs. 94

Figure 5.7: The configuration used to emulate the effects poly CD variation. The chip is divided into four quadrants, representing a fixed percentage of poly CD variation for all devices in that quadrant. 95

Figure 5.8: The clock driver consists of a series of inverters. The inverters are sized such that each successive buffer increases in size by a factor of u. The optimal value of u is equal to e=2.7182, and u=3 is taken as a good approximation..95

Figure 5.9: The D-flip flop (register) used at the outputs of the H-tree. The register is positive edge triggered, so the output latches at the rising edge of the clock. The input (V_{in}) is set to V_{DD} and the clock skew at V_{out} is due to variation in the interconnect (coming from clk) and the device variation in the register. 96

Figure 5.10: Functionality test of the latch. Data is set to high and arrives at the latch input before the clock goes high. The data is latched on the rising edge of the clock. 97

Figure 5.11: The delays of the fastest and slowest paths in the clock tree with no variation (a), copper CMP variation (b), and poly CD device variation (c). 99

Figure 5.12: Amount of oxide erosion in an array of lines depends on the bit position within the array (a). A sample profilometry trace for an interconnect array (b) shows that for the CMP process conditions in this example, lines near the edge of the array experience more metal thickness loss than those near the center....101

Figure 5.13: In an oxide CMP process, bit position within the array does not impact the ILD thickness variation. Long range interactions result in a variable ILD thickness for the entire array depending on the array location with the chip....102

Figure 5.14: The simulated delay variation as a function of bit position for a 128 bit bus. A minimum size driver in 0.25 μm technology is used to drive each line, and the pitch is varied from 1 μm to 5 μm. Minimum pitch lines near the array edge are impacted the most after copper CMP....102

Figure 5.15: The delay variation for the entire bus as a function of ILD thickness variation for underlying metal pattern densities between 0-100%. The nominal ILD thickness is targeted at 50% density....103

Figure 6.1: The 1 GHz H-tree is modified to include two stages of intermediate buffers. For clarity, the buffers are shown for only one path in the tree....107

Figure 6.2: The equivalent RC network for a symmetric H-tree from the driver to the next buffer. When a fork is encountered, the resistance of the branch is reduced by a factor of two and the capacitance doubles....108

Figure 6.3: Intermediate buffers are inserted to minimize delay when routing a signal over long distances....110

Figure 6.4: Degradation in the maximum interconnect length as a function of pitch for wires in the local, intermediate, and global tiers with worst-case variation tolerances as well as systematic variation models and tighter design tolerances....114

Figure 6.5: Propagation delay for optimally buffered interconnect (a). Delay variation with worst-case variation tolerances and tighter tolerance design (b)....116

Figure 6.6: Effects of different variation sources on signal delay with optimally buffered interconnect using worst-case tolerances (a) and tighter tolerance design (b)....117

Figure 6.7: The effect of driver size (a) and rise time (b) on crosstalk noise vs. interconnect length....120

Figure 6.8: The effect of pitch (a) and effect of metal thickness and linewidth variation (b) with minimum size drivers and input rise time of 100 psec. 121

Figure 6.9: Crosstalk noise vs. interconnect length for driver size of 10X and rise time of 300 psec with and without variation for minimum pitch (a) and 3X pitch (b). 123

Figure 6.10: Crosstalk noise for no variation, worst-case variation, and tighter tolerance design (a). Effects of different variation sources on crosstalk with worst-case tolerances (b) and tighter design tolerances (c). 125

Figure 7.1: The interconnect resistance (a), capacitance (b), and RC (c) per unit length for minimum pitch wires in the local, intermediate, and global tiers. 133

Figure 7.2: Maximum interconnect length (optimally buffered to satisfy clock frequency constraint) vs. technology generation for local (a), intermediate (b), and global interconnect (c). 135

Figure 7.3: The degradation in L_{max} for optimally buffered local (a), intermediate (b), and global (c) interconnect for different variation scaling scenarios using worst-case tolerances. 139

Figure 7.4: The degradation in L_{max} for optimally buffered local (a) and global (b) interconnect using worst-case variation tolerances and tighter tolerance design...................140

Figure 7.5: The degradation in L_{max} due to metal thickness, linewidth, device, and total variation for local (a) and global (b) interconnect at minimum pitch based on variable percent scaling scenario (ii). 142

Figure 7.6: Crosstalk noise vs. technology generation for variable pitch with optimally buffered global interconnect (a). The optimal interconnect length between adjacent buffers varies as a function of the clock frequency constraint (b). 145

Figure 7.7: Crosstalk noise vs. technology generation for minimum (a) and 3X minimum (b) pitch with no variation, worst-case variation tolerances, and tighter tolerance design.... 146

List of Tables

Table 2.1: Device Load Capacitances....48

Table 3.1: Microprocessor ILD Thickness....69

Table 3.2: ASIC ILD Thickness....70

Table 5.1: ILD Thickness Variation (%)....87

Table 5.2: Maximum Clock Skew....98

Table 6.1: Device and Power Supply Parameters and 3σ Variation....108

Table 6.2: Clock Skew in H-Tree With Interconnect, Device, and V_{DD} Variation....109

Table 6.3: Interconnect Parameters for 250 nm Technology....113

Table 6.4: Parameters for Crosstalk Noise Simulations....119

Table 6.5: Interconnect Length and Buffer Size for Optimized Crosstalk Simulations....124

Table 7.1: Scaled Device and Power Supply Parameters and 3σ Variations....129

Table 7.2: Scaled Clock Tree Parameters....129

Table 7.3: Technology Scaling Impact on Clock Skew....131

Table 7.4: Interconnect Technology Parameters....132

Table 7.5: 3σ Variation Tolerances for 250 nm and 50 nm Generations....136

Chapter 1

Introduction

Semiconductor manufacturing variation occurs when process parameters deviate from their ideal, as-designed values. Process variation has always been a key concern for manufacturability, process control, and circuit design. As technology scales, the importance of understanding variation is increasing further. Although CMOS device delay scales, interconnect delay does not [1-3]. Shrinking interconnect pitch and increasing clock frequency and chip size put greater constraints on circuit design and technology. Additionally, *variation* in the interconnect and devices results in even tighter design requirements. If process variation is not well understood, unnecessarily large design margins must be put in place to ensure that desired circuit performance specifications are met. Therefore, a key issue is to understand how much variation exists in a given design and what its impact is on circuit performance.

1.1 Sources of Variation

Variation can be categorized into temporal and spatial sources [4]. Temporal sources are time-varying and change depending on circuit operating conditions. These include effects such as switching activity, temperature variation, and reliability. Spatial effects are fixed in time and depend on physical factors such as structural variation in the chip that is based on the circuit layout, neighboring environment, and process conditions. Spatial variation sources impact the geometry of a structure and can lead to undesirable effects such as yield loss. The yield loss may be functional or parametric, causing a malfunctioning circuit or degradation in performance. Structural variation impacts both the interconnect and devices.

1.1.1 Interconnect Variation

Structural interconnect variation consists of three components: metal thickness (*T*), inter-layer dielectric (*ILD* or *H*) thickness, and linewidth (*W* or *LW*) as shown in Fig. 1.1. Additional geometric effects such as sidewall slope or surface and edge roughness may also be of concern, but are not considered in this work. Note that linespace (*S* or *LS*) is not an independent parameter since a variation in linewidth automatically causes a change in the linespace. Variation in the interconnect results in a change in its electrical properties, including the resistance (*R*), capacitance (*C*), and inductance (*L*). These electrical parameter variations directly affect the performance of the circuit. The critical paths often contain long wires, and a good description of the interconnect geometry variation is needed for accurate circuit simulation.

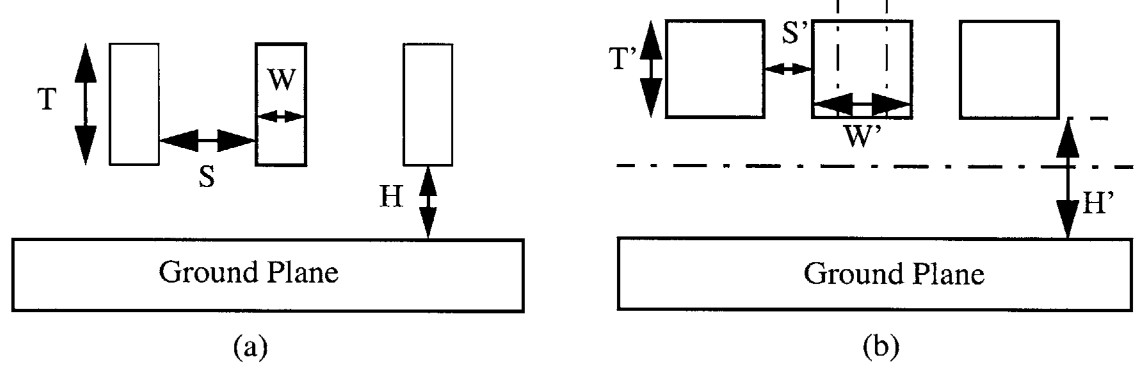


Figure 1.1: Cross-section of parallel interconnect lines above a ground plane. The figure on the left (a) shows the ideal case, and the figure on the right (b) shows some of the different types of variation that can exist in the interconnect.

1.1.2 Device Variation

Structural variation in the devices includes gate length (L_{gate}), gate width (W_{gate}), and gate oxide thickness (t_{ox}). Among the other sources are variation in the drain and source active areas as well as variations in the doping profiles. All of these types of variation

change the device properties and affect circuit performance (see Fig. 1.2). The most important sources of device variation are L_{gate}, t_{ox}, and V_t (threshold voltage). Since the ratio of W_{gate}/L_{gate} determines the drain current of a CMOS transistor, if W_{gate} is much larger than L_{gate}, variation in W_{gate} is usually not considered.

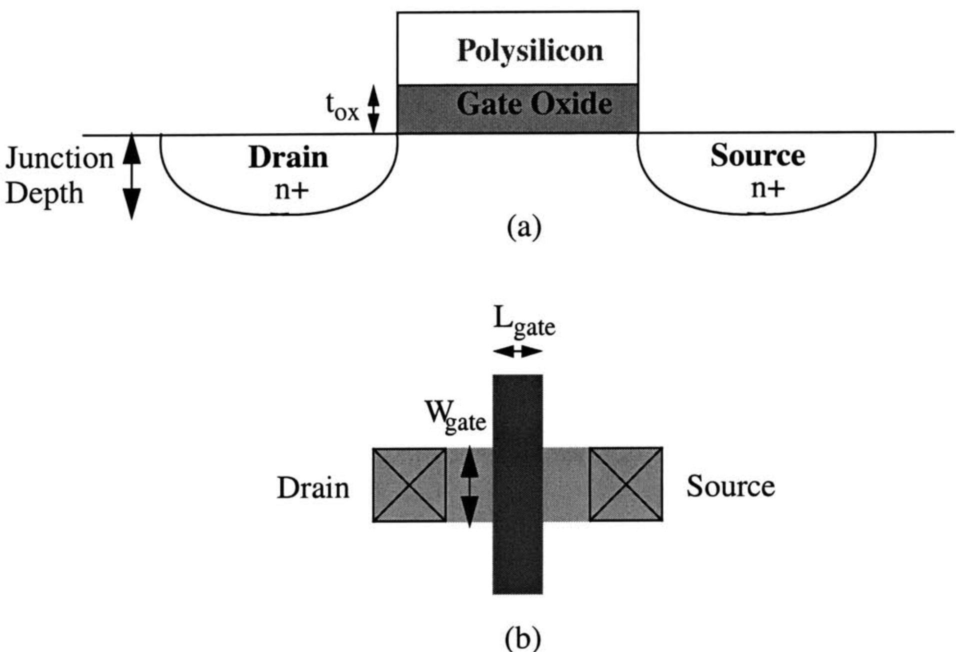


Figure 1.2: Cross-section (a) and top view (b) of an NMOS device on a silicon wafer. Structural variations in the polysilicon and gate oxide impact the performance of the device in addition to variations in the active areas or doping profiles.

1.2 Motivation

Several works have previously considered the impact of variation on circuit performance. Most earlier studies have focused specifically on device variations, e.g. [5,6], since interconnect has become an important issue only very recently. Conventional statistical analysis techniques typically assume that circuit parameters are independent random variables with a Gaussian distribution. Let $X = [X_1, X_2,..., X_N]$ represent the input vector of normally distributed random variables with mean μ_i and standard deviation σ_i for

$i=1,2,...N$. The input vector typically consists of either geometric parameters (e.g. LW, T, L_{gate}) or electrical parameters (e.g. R, C) in the interconnect or devices [7-9]. The output vector $Y = [Y_1, Y_2,..., Y_M]$ is a function of X and contains circuit performance variables such as signal delay, clock skew, or crosstalk noise.

The simplest approach to guarantee that performance specifications will be met is the skew-corner model. Here, the objective is to pick the corners for input variables X_i such that an acceptable yield or performance criteria is met. However, if the inputs X_i are correlated or if an output variable Y_i is a non-linear function of X_i, it may be difficult to pick the true corners. Consider the bivariate example shown in Fig. 1.3, with inputs X_1 and X_2 and output Y. One approach is to pick the corners based on "worst-case limits" (the four corners of the rectangle in Fig. 1.3) by taking values that are several standard deviations away from the mean ($\mu_i \pm k\sigma_i$). Typically, k=3 is used to obtain a 99.73% confidence level. However, this technique can result in overly pessimistic estimates of the output performance: if the inputs are correlated, all the samples may lie within the ellipse shown in Fig. 1.3. In this case, the true input variable tolerances are much tighter than those obtained by the minimum and maximum limits. Since there is no output performance distribution, however, it is hard to determine what more realistic tolerances (corners) are if the function $Y = f(X_1, X_2)$ is not known.

If the function $Y = f(X)$ can be analytically or numerically computed, a useful approach is to run Monte-Carlo simulations [7-12]. With this technique, random samples of each input variable X_i are taken. The outputs are computed for each set of inputs over several trials (hundreds or thousands) and a distribution results. The advantage is that a realistic distribution is obtained for the output. Using this distribution, more realistic corners can be obtained for the inputs by setting a confidence level on the output. The disadvantage is that

simulation time increases significantly. Fig. 1.4 qualitatively shows how the Monte-Carlo method can be used to obtain the input corners.

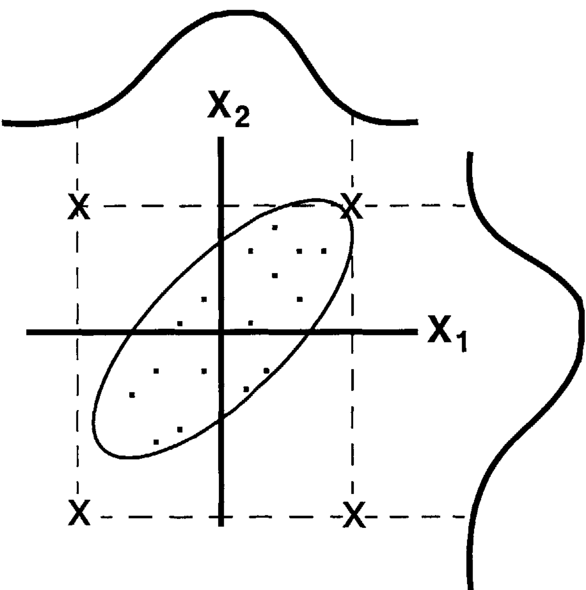


Figure 1.3: The input variables X_1 and X_2 are randomly distributed. If the inputs are correlated, the samples may lie within the region bounded by the ellipse. In this case, using minimum and maximum limits can result in pessimistic corners.

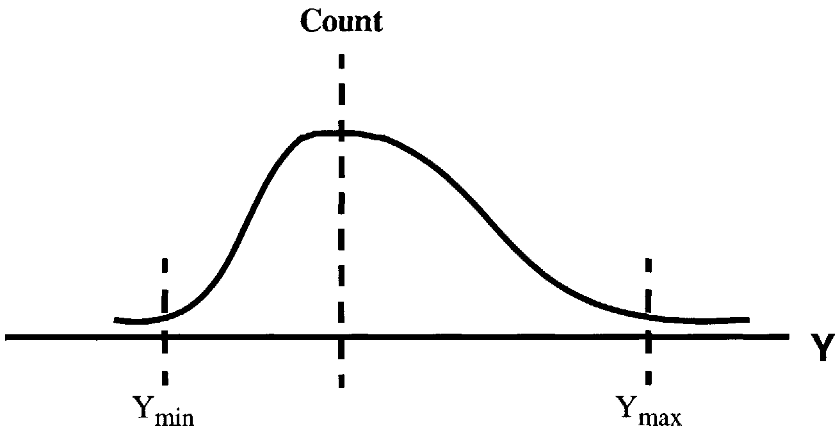


Figure 1.4: The output distribution is obtained by taking random samples of the inputs X_i and performing several simulations (typically thousands). The input corners can be obtained by determining the values of the inputs such that the output is inside the specified confidence interval (i.e. find bounds for X_i such that $Y_{min}<Y<Y_{max}$).

Most recent works that have analyzed the impact of variation on circuit performance continue to use approaches that are based on the statistical performance analysis techniques described above. Although these methods result in narrower design margins over the skew-corner approach by using random distributions for the variation sources, they fail to account for *systematic* variation effects which may be a large fraction of the total variation. The next section discusses these different variation components.

1.2.1 Systematic vs. Random Variation

While the assumption of randomly varying parameters is a good one for *most* device variation sources, it is not true in general for interconnect parameters. Until recently, the Monte-Carlo analysis method has been adequate since the focus has been mainly on devices. However, with technology scaling, the increased impact of interconnect on signal delay has placed a greater importance on interconnect variation. Studies have shown that a large fraction of the variation in the interconnect may be a function of the layout characteristics [13, 14]. Rather than describing variation as a purely random source, some parameters may have a large systematic or deterministic component. While previous studies have placed an emphasis on lot-to-lot, wafer-to-wafer, and within wafer variations (see Fig. 1.5), it is the intra-die (also known as within-die or across chip) variation that has recently become a very real concern. For any parameter, given that a significant portion of the total variation is systematic, *deterministic* models for the systematic component of variation can be used to predict much of the variation. Rather than using worst-case corners or Monte-Carlo methods to bound the variation, models can be used for the systematically varying components, thereby reducing overall uncertainty. The variation can be separated into components [15, 16] using a "statistical metrology" framework.

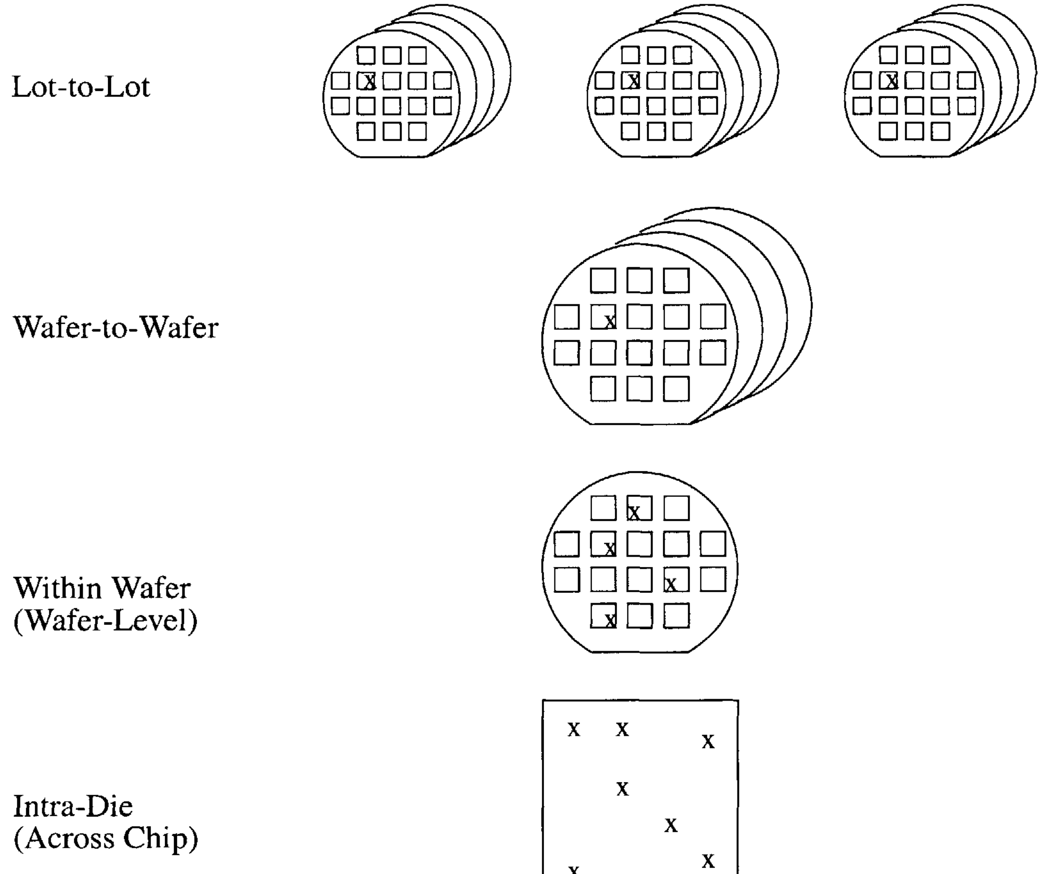


Figure 1.5: Variation can be decomposed into different components, including lot-to-lot, wafer-to-wafer, within wafer, and intra-die. The intra-die component can be a large component of the total variation. For example, the most significant source of interconnect ILD thickness variation is the intra-die component.

The techniques described in [17] provide a methodology for variation decomposition (see Fig. 1.6). Raw data from a single wafer is taken as input and in each stage of the process a component of the systematic variation is modeled. The residuals are fed to the next stage and an additive model is used. The sequence progresses from the wafer-level estimation to intra-die and the wafer-die cross term estimation. The residuals left at the end of the process are assumed to be random sources that cannot be explained. Different types of estimators are applied to extract the various components. The wafer-level variation is a smoothly varying component and is generally attributed to process and equipment factors.

Downsampled moving averaging, splines, or regression based estimators may be used here. The die-level component usually contains high frequency content because of the short range effects in the layout pattern. For estimating the intra-die variation, a Fast Fourier Transform (FFT) is generally used since the die pattern is periodic throughout the wafer. Finally, the wafer-die cross term can be calculated using splines or FFT analysis.

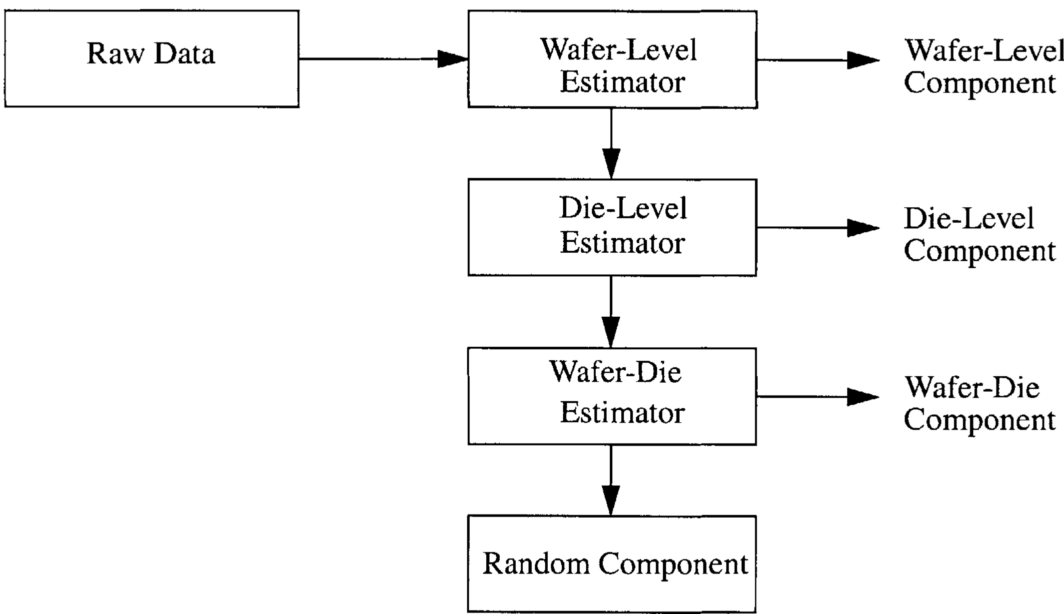
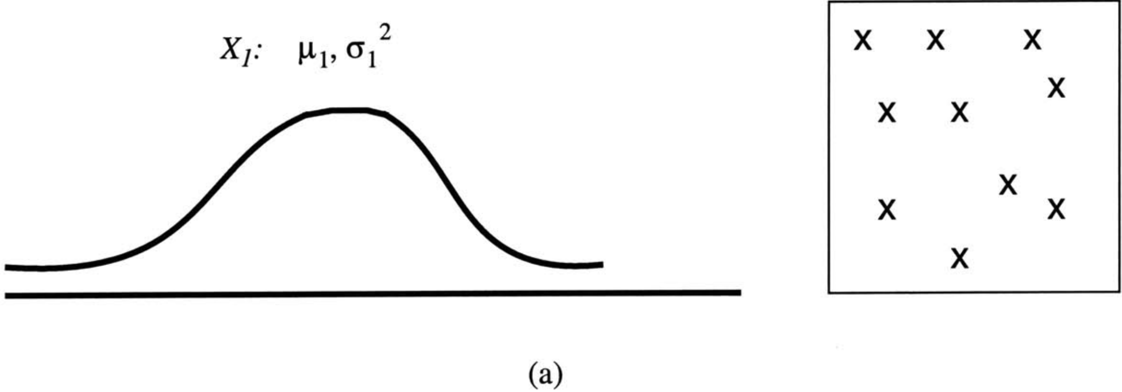


Figure 1.6: The variation is separated into components using a series of estimators. After each stage, a variation component is extracted and the residuals are fed to another estimator for extracting the next component [17].

Fig. 1.7 qualitatively shows how the systematic variation modeling approach can help reduce design uncertainty. In Fig.1.7 (a), samples from a single die are taken to obtain a distribution of input variable X_I. In the Monte-Carlo analysis, it is assumed that this distribution is random. In reality, however, the total variation may be a function of the spatial location within the die (see Fig 1.7 (b)). Here, each point within the die has a mean value of X_I that is dependent on the layout pattern with a much narrower distribution than that obtained by sampling various points within the die. The distribution may be obtained by

sampling the same within-die location across multiple die within a wafer and across several wafers. If this is the case, the within-die component is considered to be systematic and the other sources are assumed random.

Random Sampling:



Variation Decomposition:

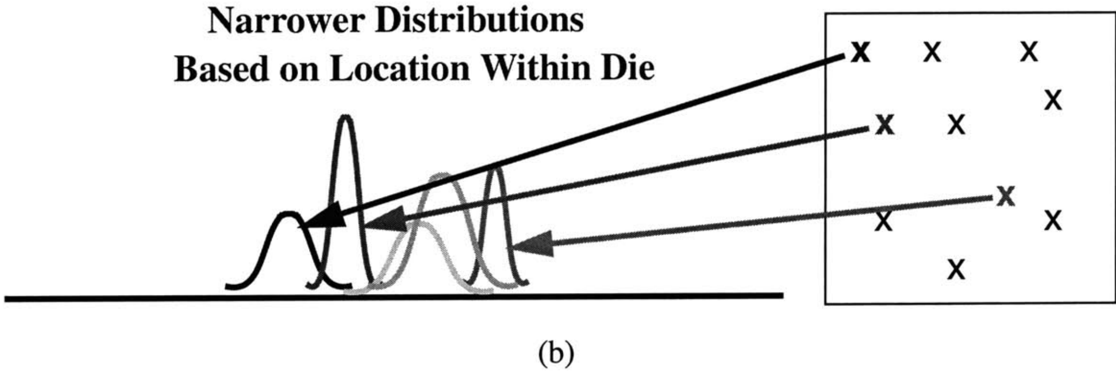


Figure 1.7: Random sampling within a single die is used to determine the statistics for input variable X_1 (a). Each structure within the die is sampled across multiple die within a given wafer for several wafers (b).

In this thesis, we explore several sources of systematic variation. A major source of variation results from planarization of the interconnect. This includes the inter-layer dielectric (ILD) or metal thickness variation. We now describe how pattern dependent interconnect variation results from chemical mechanical polishing (CMP).

1.2.2 Systematic Variation in Interconnect CMP Processes

CMP is a commonly used technique to planarize the interconnect or ILD between adjacent metal layers [18]. Fig. 1.8 shows a rotary polishing tool used for CMP. The wafer is held on a carrier while the platen holds a porous pad that is used for polishing. Both the platen and carrier rotate as a slurry material is fed to the polishing pad. The slurry consists of an abrasive material that provides the chemical and mechanical action.

Although CMP provides good local planarization, global non-uniformity still exists after CMP. Fig. 1.9 shows a short flow metal etch process used with aluminum interconnect. A blanket layer of oxide is first deposited. This is followed by metal deposition and patterning. Oxide is deposited in the next step and provides the insulator (or inter-layer dielectric) that exists between two adjacent layers of metal. The goal is to planarize this material to obtain a smooth surface before depositing the next layer of aluminum.

The amount of material removed depends on several process conditions such as the pad pressure, slurry type, polish rate, polish time, and the *layout pattern*. The underlying metal pattern is a very important factor. Fig. 1.10 shows the post-CMP inter-layer dielectric thickness variation for a single die [19]. We see that the across chip (intra-die) variation is the most significant component compared with the wafer-level, wafer-die cross term, and random components.

The main cause of ILD thickness variation in an oxide CMP process is pattern density [20]. The pattern density is defined as the amount of metal in a given region divided by the

total area of that region. The range over which density is calculated depends on the CMP process and is known as the CMP process planarization length or interaction distance. Physically, this depends on how the pad conforms to the patterned features. Fig. 1.11 shows the effect of pattern density on the post CMP ILD thickness in an oxide CMP process. There is one region of high density (several closely spaced metal lines) and another region of low density (metal lines spaced further apart). With conformal deposition, the oxide step height follows the shape of the underlying metal pattern. In dense areas, there is a larger amount of oxide deposited than in sparse areas. The reasoning behind CMP ILD thickness variation is based on the volume of oxide removal within a given area: a fixed volume of oxide will be removed for a specified polish time. Since all areas on a wafer must polish for the same amount of time, the region of low metal density will polish the oxide at a faster rate than a region of high density. The oxide is typically polished until all oxide features have been removed and a smooth surface exists over the entire wafer. After a region of low density has been completely polished, that area starts to polish at the blanket oxide removal rate. This starts thinning down the oxide in a low density region even though high density areas have not been completely cleared. The result is that after CMP, there is more oxide remaining above the dense areas since there is more oxide at the start of the CMP process.

CMP is a complex process and we have only highlighted the basic ideas here. There are several good resources on CMP and a description can be found in [18]. In Chapter 3, we will describe the differences between CMP processes (metal vs. oxide) in more detail, including pattern density calculation and the effects of different planarization lengths. We will also provide the models used to simulate the effects of CMP in creating interconnect variation.

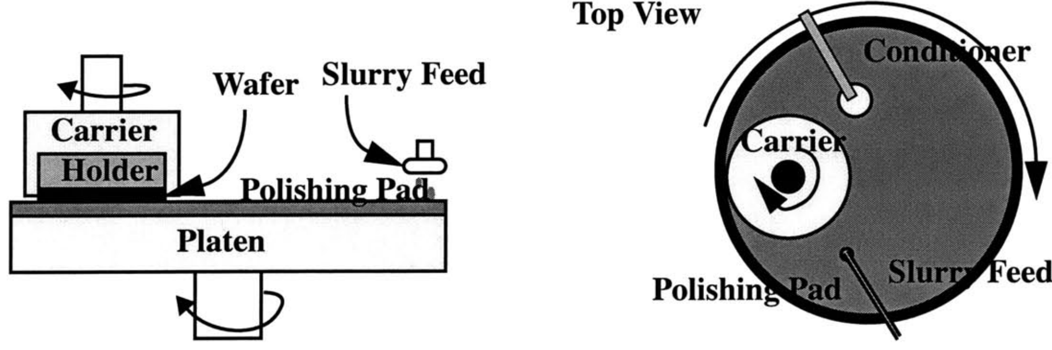


Figure 1.8: A rotary CMP tool. Both chemical and mechanical action is used in polishing the wafer. The carrier and platen rotate as a slurry chemical is fed to the polishing pad.

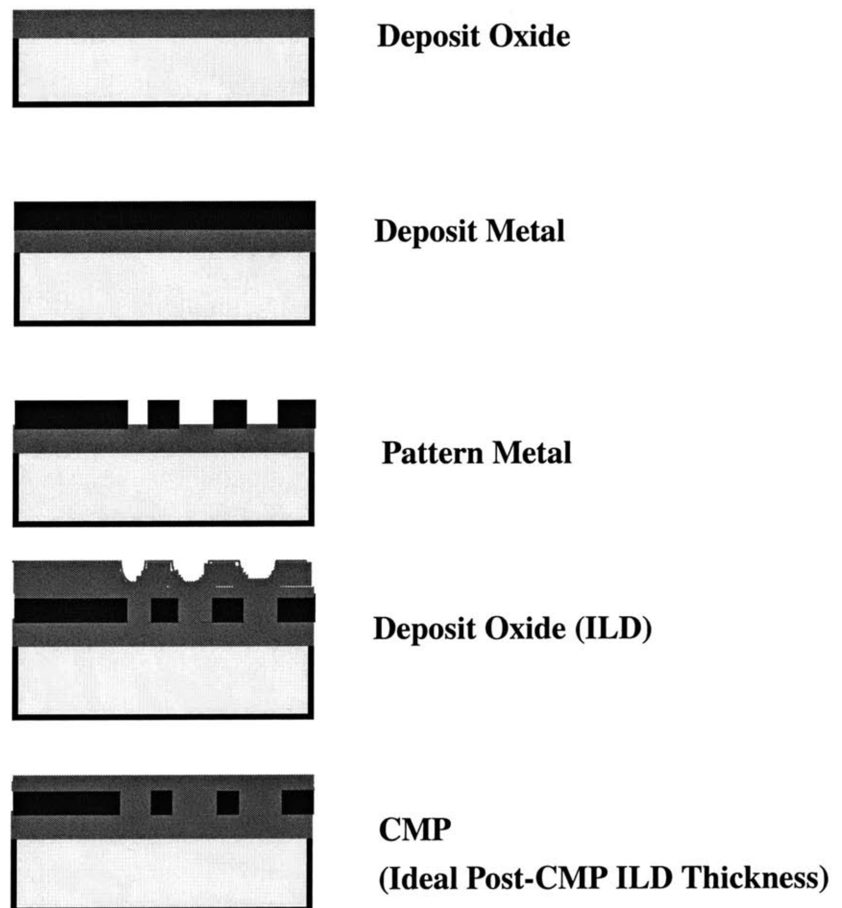


Figure 1.9: A short flow process where the metal is patterned, followed by oxide deposition and CMP. In the ideal case, the dielectric is polished so that a smooth and flat surface results after CMP.

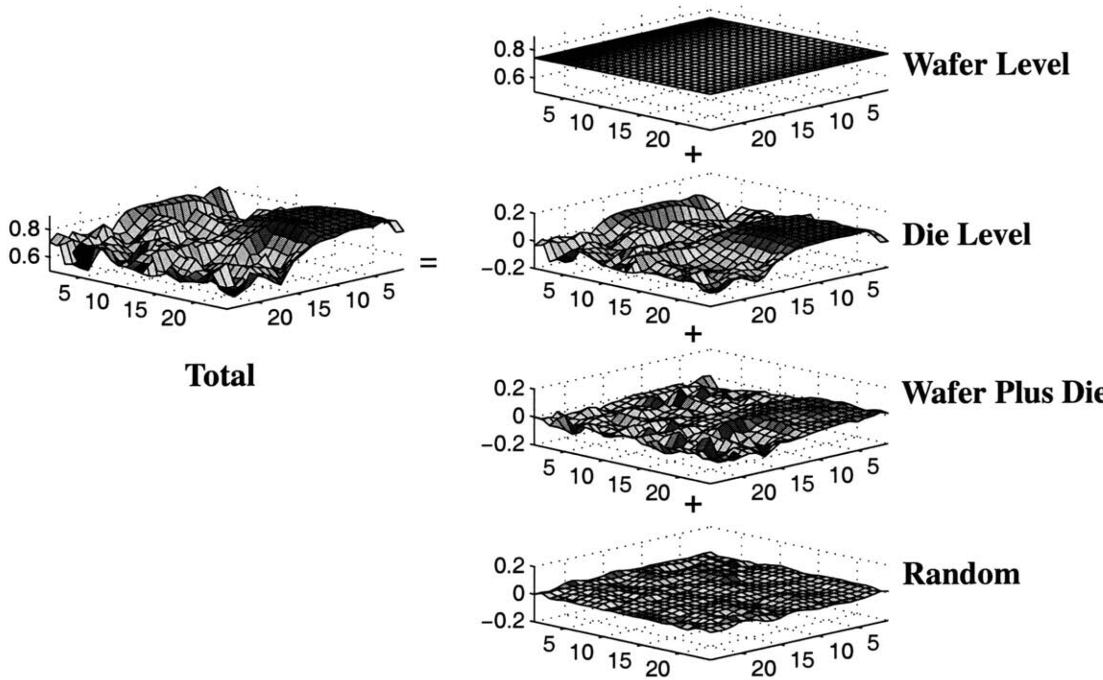


Figure 1.10: The ILD thickness variation across any individual die can be separated into die-level, wafer-level, wafer plus die, and random components. The greatest amount of non-uniformity is in the die-level component, indicating the importance of pattern dependence.

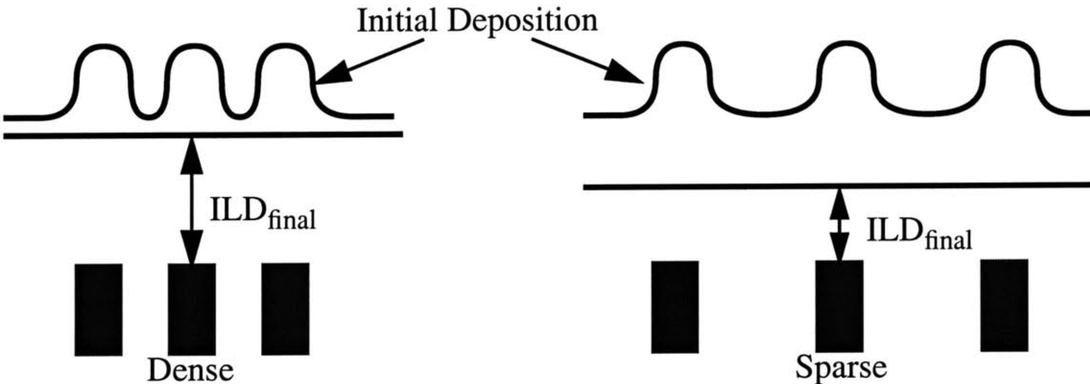


Figure 1.11: Post CMP ILD thickness is a strong function of underlying metal pattern density. With conformal deposition, sparse regions polish faster than dense regions, resulting in a greater amount of oxide above dense areas.

1.2.3 Interconnect Variation Impact on Circuit Performance

Interconnect variation has a direct impact on circuit performance. This variation can impact three important metrics in high performance digital circuits such as microprocessors. The first of these is signal path delay (see Fig. 1.12). Long wires are used for across chip communication, and critical paths are usually limited by the interconnect [3]. Variation in the interconnect affects its electrical properties (resistance, capacitance, inductance) and may increase delay. If the signal does not arrive at the output in the required time, circuit malfunction can occur. A long wire limits the amount of time a signal has to arrive at the output, and when variation effects are included tighter design guidelines may need to be imposed for proper circuit operation.

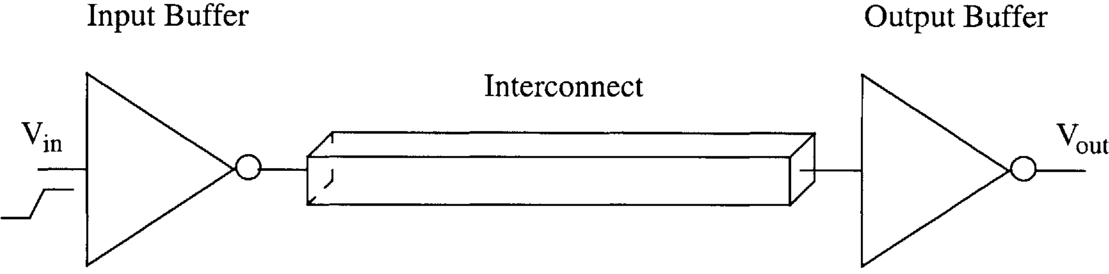


Figure 1.12: Interconnect delay is generally a large fraction of the total signal delay for critical paths. Variations in the interconnect impact its electrical properties and can limit circuit performance.

Another important circuit metric is clock skew. This refers to matching the signal delays at the outputs of the tree. Since the clock is a crucial element of synchronous design, minimizing clock skew is a top priority for any circuit. One way to minimize the skew is to make the tree symmetric. This can be accomplished with an H-tree configuration (see Fig. 1.13). Wires in the tree are tapered for impedance matching so that when a fork is encountered, the widths of the branches in the next level are reduced by a factor of

two. The most important consideration for a good clock design is that the *difference* in signal arrival times at the outputs is very small, and absolute delay does not matter. However, with interconnect variation, even a symmetric design can be susceptible to clock skew.

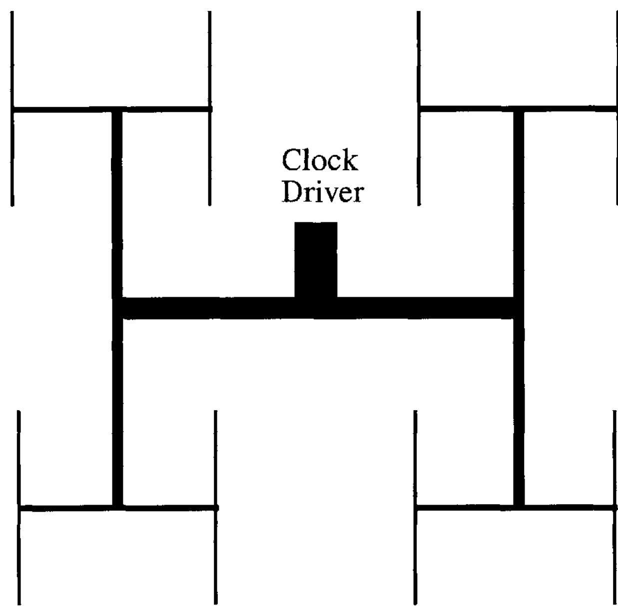


Figure 1.13: An H-tree used for clock distribution. Wires in the tree are tapered for impedance matching [2]. When a fork is encountered, the width of the next branch is reduced by a factor of two.

The third important performance metric is crosstalk noise. This occurs when a neighboring wire unintentionally influences the behavior of another wire (see Fig. 1.14). Small fluctuations of a signal state usually do not affect the performance of a circuit. However, excessive noise can cause a signal to change its state (e.g. high to low) and result in circuit malfunction. This depends on how much noise margin is available and is a big issue in low swing circuits such as sense amplifiers. Since crosstalk is affected by lateral coupling between adjacent wires, linewidth variation may increase the noise to unacceptable levels.

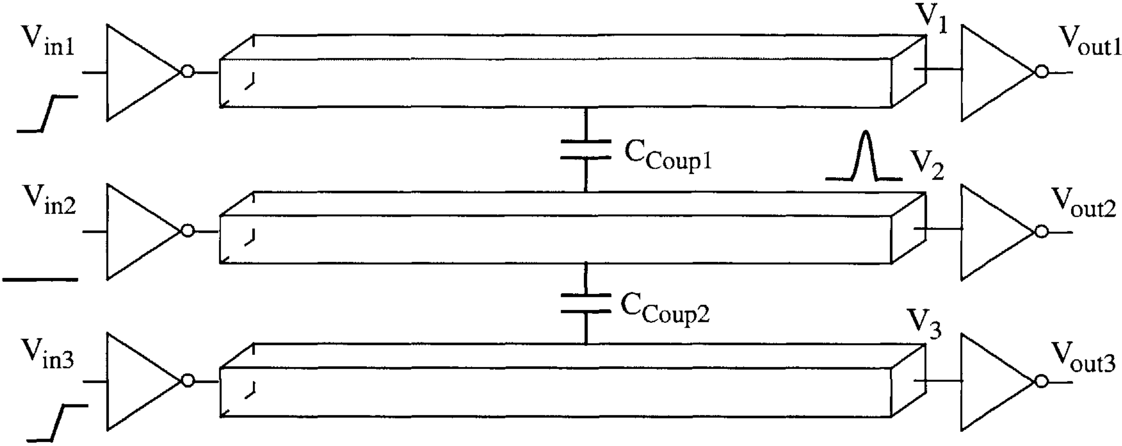


Figure 1.14: The input signal V_{in2} is quiescent, but its neighbors are switching from low to high. Due to the lateral coupling between lines, crosstalk noise causes a spike at V_2.

1.2.4 Technology Scaling Impact

Although variations in the interconnect have a direct impact on circuit performance, these effects become even more important as technology scales. Fig. 1.15 shows a three-tier wiring network configuration that is used in most high performance designs. Wires in the local tier are typically used for wiring over short distances, while those in the intermediate and global tiers are used for communication over long distances. Additionally, the global tier is also used for the power grid and clock distribution.

The impact of technology scaling is best seen by considering the interconnect delay as a function of technology. We start with the SIA Roadmap [1] to obtain the projected interconnect parameters and global clock frequencies as technology scales from the 250 nm generation to the 50 nm generation. In Fig. 1.16 (a), we plot the maximum interconnect length that can be used for routing assuming that almost all of the delay is due to the interconnect. This is an optimistic case since realistically there will be additional stages of logic in between. However, repeaters (intermediate buffers) may be inserted to reduce the delay. In Fig.1.16 (b), the maximum interconnect length is plotted assuming that optimal

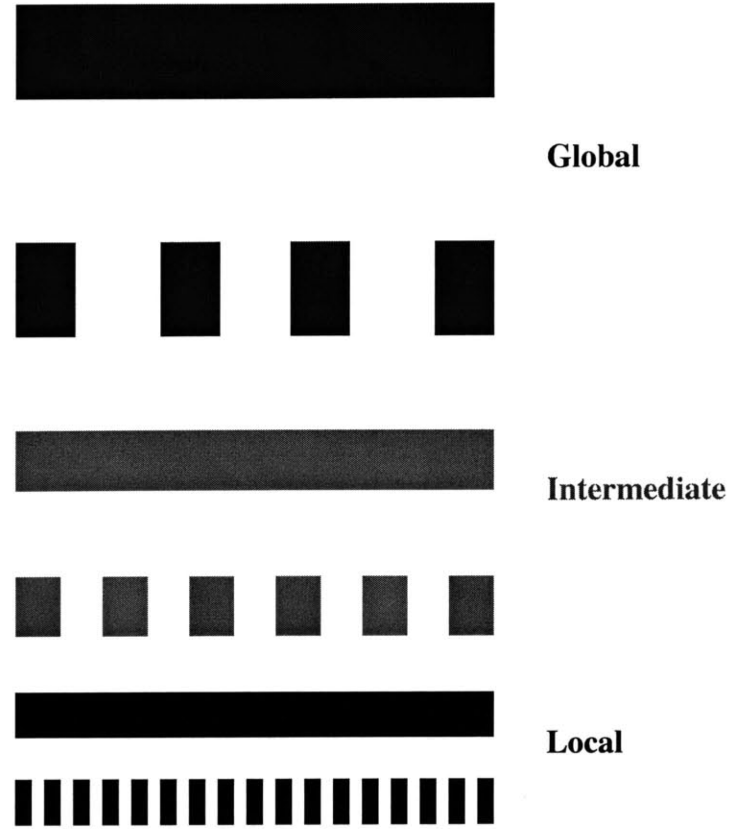


Figure 1.15: A three-tier interconnect wiring scheme typically found in high performance circuit designs. Each tier contains two or more wiring levels, including alternating levels for horizontal and vertical routing. Wires in the intermediate and global tiers are used for long distance routing. Power supply and clock lines generally use global wires.

buffer insertion is used. The maximum interconnect length is plotted for minimum pitch (*LW*+*LS*) wires in the global, intermediate, and local wiring tiers. Additionally, since the maximum distance over which one is likely to communicate is two times the chip side length (assuming only horizontal and vertical wiring), this value is also plotted. Note that although the chip size is expected to increase as technology scales toward the 50 nm generation, the maximum wire length available to meet the global clock frequency requirement decreases. We see that by the 100 nm generation, even in the ideal case where the path delay is due to just the interconnect and repeaters, global wires will not be able to route the longest paths in one clock cycle. For these wires, additional pipeline stages must

be added, and the chip complexity will increase. However, wires that are just short enough to meet the timing requirements will be the most critical. For these cases, accurate interconnect modeling is important. A good understanding of the variation allows for a more aggressive design without sacrificing performance or increasing design complexity.

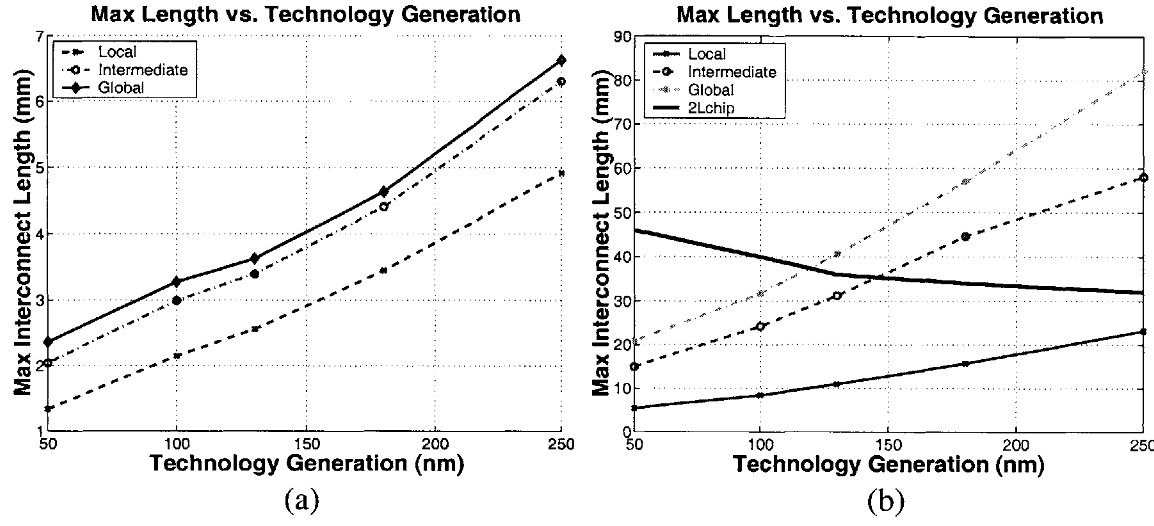


Figure 1.16: The maximum distance that can be routed using minimum pitch local, intermediate, and global interconnect to meet the global clock frequency constraint for different technology generations without intermediate buffers (a) and with buffer insertion (b). By the 100 nm generation, the longest global wires will require more than one clock cycle even with intermediate buffers.

1.3 Thesis Goals

In this chapter, we have provided an overview of process variation in semiconductor manufacturing. We have shown that variation can be categorized into different types, including spatial vs. temporal, device vs. interconnect, and systematic vs. random. We have also pointed out that a large fraction of the variation may be due to spatial systematic effects, particularly in the interconnect. This variation can negatively impact circuit performance and produce designs that do not meet specifications, resulting in functional and parametric yield loss. Additionally, as technology scales, these problems are likely to become worse due to tighter design requirements.

Although variation can have a detrimental impact, a lack of understanding about the nature of variation can be even worse. Most design guidelines assume that semiconductor process variation is random. However, from our discussion earlier in the chapter, we know that this is not always true. We have shown that CMP ILD thickness variation has a large systematic within-die component that is based on the layout pattern. Given that most of the variation is systematic for a given parameter, we can model the variation based on spatial effects such as the location and geometric pattern. Variation modeling can improve performance regardless of whether or not variation is a concern in a given design. If the amount of variation is unacceptable, pattern dependent models can determine what parts of the design need to be corrected. Even if variation is not a big concern, performance can be improved by tightening the design margin for one parameter and allowing greater flexibility for another parameter. This can allow the designer to avoid any unnecessary overdesign. We note this important observation and propose a new methodology for simulating the impact of systematic variation on circuit performance.

The main goals and contributions of this thesis are outlined here as follows:

- Provide a methodology for simulating the impact of systematic interconnect and device variation on circuit performance.
- Implementation of methodology within a CAD framework to be compatible with existing circuit layout, net extraction, and performance simulation tools.
- Industrial case studies that utilize the new methodology and assess the impact of different sources of systematic process variation.
- Demonstrate the relative importance of random vs. systematic variation effects.
- Study the effects of technology scaling to demonstrate the increased importance of including systematic variation models for circuit simulation.

1.4 Thesis Organization

This thesis is organized into eight chapters that illustrate the goals outlined above. Chapter 2 starts with a short description of the different types of interconnect models that are used with circuit simulation tools. It explains the dependence of the interconnect and device parameters on different circuit performance metrics by reviewing existing analytical models. Chapter 3 discusses systematic process variation models for different interconnect and device variation sources. In Chapter 4, we describe our new methodology for simulating the impact of systematic variation and its implementation to interface with existing CAD tools. Chapter 5 applies our methodology to study the impact of variation on high performance industrial designs. These case studies include signal delay variation and clock skew in two different 1 GHz microprocessors designed in the 250 nm and 180 nm generations using aluminum and copper interconnect. In Chapter 6, we include the effects of both systematic and random variations in the interconnect and devices to get a better perspective on the impact of variation. Additionally, we also study the impact of variation on crosstalk noise. In Chapter 7, we study the effects of technology scaling based on the SIA roadmap projections and different variation scaling scenarios. Finally, Chapter 8 concludes the thesis and provides directions for future work in this area.

Chapter 2

Circuit Performance Metrics

Digital circuit performance metrics consist of several aspects of circuit design. In this thesis we explore variation impact on three of these metrics, all of which are important in interconnect dominant circuits. The first is signal delay, which requires that the signal arrival times meet the specified timing constraints. The second is clock skew, which refers to the difference in the arrival times among signals. In synchronous design, minimizing clock skew is essential for proper circuit operation without unnecessary over-design. Finally, the third issue is signal integrity or crosstalk noise, which may increase due to variation in the interconnect. Signal integrity requires proper circuit operation with minimal interference from neighboring signals. In this chapter, we review the basic concepts and fundamental equations to show how variation affects delay, skew, and crosstalk noise in digital integrated circuits. We start with a discussion of interconnect modeling and then describe analytical models for each of the three metrics.

2.1 Interconnect Modeling

The interconnect delay is a function of the transistor on-resistance, device load capacitance, and the interconnect load. The interconnect load is distributed and is modeled as a function of the wire resistance and capacitance (and inductance for high enough frequencies) which depend on the wire geometry. At low frequencies, the interconnect is modeled as a distributed RC network [2]. To represent the distributed nature of the interconnect, it is broken down into smaller lumped sections for simulation. The possible representations include the pi, T, and ladder networks. We consider the RC network shown in Fig. 2.1, where the line is divided into N sections. The simulation accuracy increases with increas-

ing N. If the signal rise time is too fast or the wire is very long, the inductance must also be included and an RLC network must be used to account for transmission line effects (see Fig. 2.2).

Figure 2.1: The pi (a), T (b), and ladder (c) representations of the interconnect. A long wire is broken down into N sections for circuit simulation to account for the distributed effect.

Figure 2.2: A distributed RLC representation of the interconnect network. The inductance must be included at high frequencies.

2.1.1 Capacitance Calculation

The capacitance can be computed using either a 2D or 3D capacitance solver such as Raphael [21] or FASTCAP [22] or closed form models, depending on the level of accuracy desired. The simplest equation is for the parallel plate capacitance and is given by

$$C = \frac{\varepsilon A}{D} \tag{2.1}$$

where

$\varepsilon = \varepsilon_0 \varepsilon_r$ is the dielectric constant,

A = area of overlap between the two conductors,

D = distance of separation between conductors.

For interconnect in integrated circuits, however, additional effects must also be included. Since on chip interconnects typically have high aspect ratios (height to width), a large part of the capacitance comes from lateral coupling between adjacent wires on the same metal level. The total capacitance includes overlap, lateral, and fringing effects (see Fig. 2.3 (a)). For delay analysis, the total capacitance may be lumped into a single value. This can be approximated using closed form models such as those given in [23]. The capacitance per unit length l for a wire with two adjacent neighbor lines above a ground plane is given as

$$C = \varepsilon\left(1.15\left(\frac{W}{H}\right) + 2.8\left(\frac{T}{H}\right)^{0.222} + 2\left(0.03\left(\frac{W}{H}\right) + 0.83\left(\frac{T}{H}\right) - 0.07\left(\frac{T}{H}\right)^{0.222}\right)\left(\frac{S}{H}\right)^{-1.34}\right) \tag{2.2}$$

where W, S, T, and H are the geometric parameters as defined in section 1.1.

If the neighbor lines are also switching the capacitance equations above must be modified to account for the correct inter-layer coupling capacitance. If a neighbor line is switching in the opposite direction, the effective lateral coupling capacitance doubles. If the neighbor is switching in the same direction, there is no lateral coupling and this capacitance is equal to zero.

For crosstalk noise calculations, the total capacitance cannot be lumped together (see Fig 2.3 (b)). The line-to-line and line-to-ground capacitances must be separated since crosstalk depends on the ratio of intra-layer coupling capacitance to total capacitance. In

this case the coupling capacitance can be computed using formulas proposed by Sakurai [24]. For the case of two adjacent neighbors above a ground plane, the intra-layer capacitance per unit length is given as

$$C = \varepsilon\left(1.93\left(\frac{T}{H}\right)^{1.1} + 1.14\left(\frac{W}{H}\right)^{0.31}\right)\left(\frac{S}{H} + 0.51\right)^{-1.45} \qquad (2.3)$$

The line-to-ground capacitance can then be found by subtracting Eq. 2.3 from Eq. 2.2. Closed form equations for other cases such as one neighbor line or two ground planes are also available and are given in [24, 25].

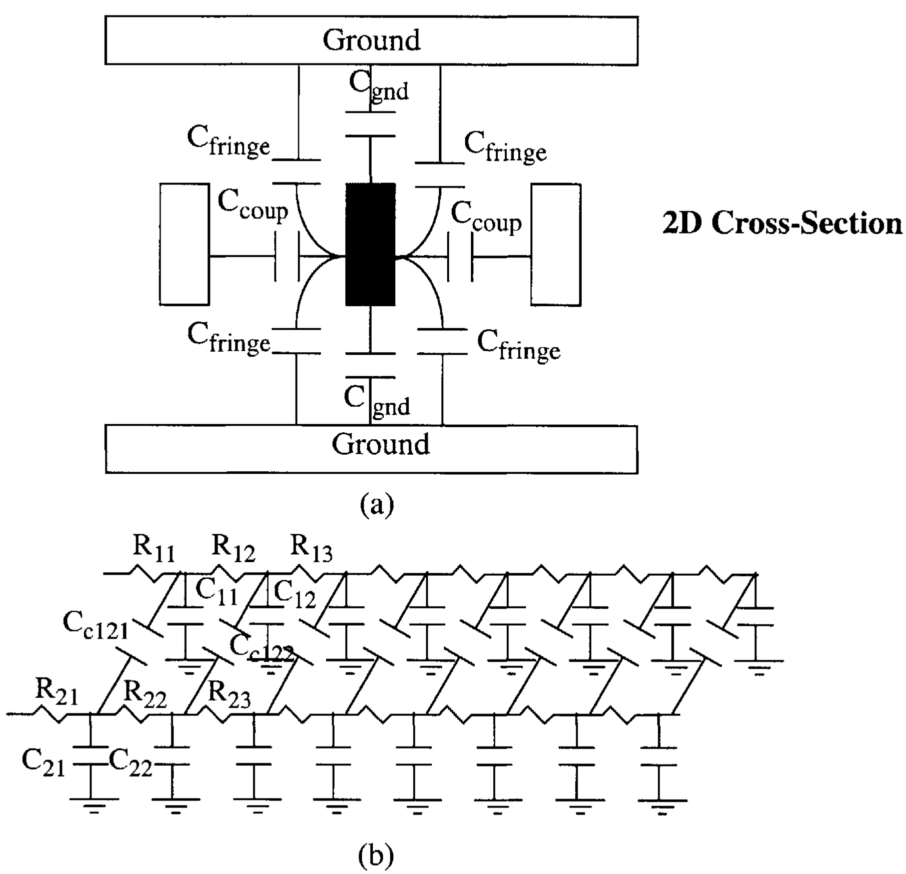


Figure 2.3: The interconnect capacitance is a function of wire geometry. It includes ground capacitance, lateral coupling, and fringing effects (a). The interconnect can be modeled as a series of distributed RC sections that include separate line-to-ground and line-to-line coupling capacitances (b).

2.1.2 Resistance Calculation

The resistance per unit length for an interconnect is calculated as

$$R = \frac{\rho}{TW} \tag{2.4}$$

where ρ is the resistivity of the metal [26]. At high frequencies, the current pushes towards the surface of a conductor and the resistance increases. This is known as the skin effect. The skin depth is given as

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}} \tag{2.5}$$

where

f = signal frequency,

μ = magnetic permeability of material,

σ = conductivity of interconnect ($1/\rho$).

When the skin depth becomes less than the dimensions of the interconnect, the resistance increases and the skin effect must be considered.

2.2 Signal Delay

The signal delay is a function of the driver resistance, load capacitance, and interconnect RC. Consider the example given in Fig. 2.4. It consists of an input buffer driving a load containing a long wire and an output buffer. The signal delay (to first order) is given by Bakoglu [2] as

$$T_d = 0.4 R_{int} C_{int} + 0.7 (R_{tr} C_{int} + R_{tr} C_L + R_{int} C_L) \tag{2.6}$$

where

R_{tr} = transistor on-resistance,

R_{int} = total (lumped) interconnect resistance,

C_{int} = total (lumped) interconnect capacitance,

C_L = device load capacitance.

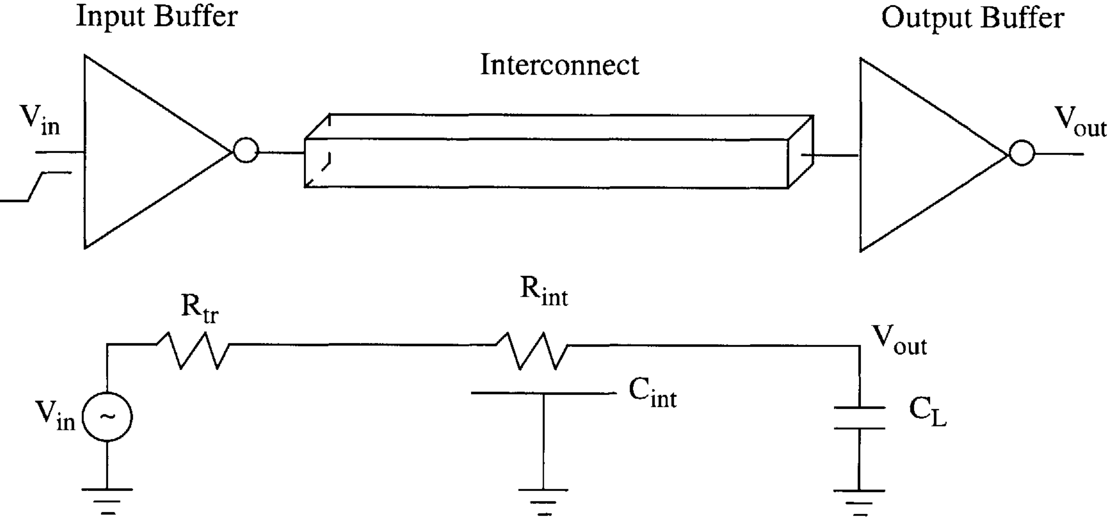


Figure 2.4: The signal delay is modeled using the circuit shown above. The total path delay includes the effects of the driver resistance, device load capacitance, and the distributed RC interconnect load.

The input buffer (driver) is modeled as a switch with a fixed resistance (see Fig. 2.5). We assume that the NMOS transistor turns on and the PMOS turns off immediately in a high to low output transition and vice-versa. The resistance of the driver is estimated by averaging the drain currents at the endpoints of the low-to-high or high-to-low transitions. The resistance is then approximated as the difference in the endpoint voltages divided by the average drain current. The drain current of the NMOS transistor in saturation is given as

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{Tn})^2 \quad (2.7)$$

and the current in the linear region is given as

$$I_D = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{Tn})(V_{DS}) - \frac{V_{DS}^2}{2} \quad (2.8)$$

where

L = transistor gate length,

W = NMOS transistor gate width,

μ_n = mobility of NMOS transistor,

C_{ox} = NMOS gate capacitance per unit area.

The endpoint voltages are V_{DD} and GND in a full swing circuit. For an NMOS transistor, the on-resistance can be approximated as

$$R_{trn} = \frac{L}{W(\mu_n C_{ox}(V_{DD} - V_{Tn}))}. \quad (2.9)$$

The device load C_L consists of the input capacitance of the output buffer, which includes the gate and drain capacitances of the NMOS and PMOS transistors (see Fig. 2.6). The different components of the load capacitance are taken from Rabaey [3] and summarized in Table 2.1. The junction (*CJ*), sidewall (*CJSW*), and overlap (*CGDO*) capacitances are obtained from the Spice models for the devices. The drain and source areas and perimeters are referred to as *AD* and *PD*.

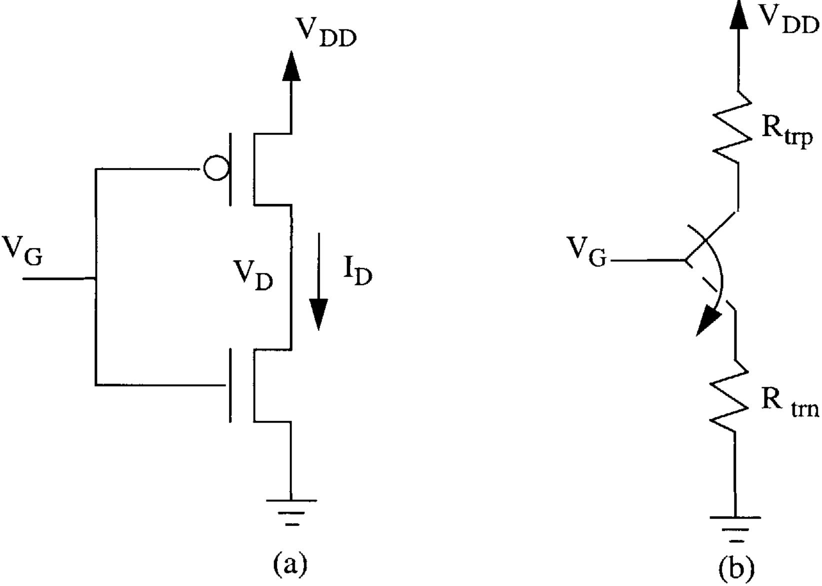


Figure 2.5: The CMOS inverter (a) is modeled as a switch with a finite on-resistance (b) for computing the signal delay. The NMOS resistance is used for a low to high input transition and the PMOS resistance is used for a high to low input transition.

Table 2.1: Device Load Capacitances

Capacitance	Expression
C_{gdn}	2 CGD0 W_n
C_{gdp}	2 CGD0 W_p
C_{dbn}	K_{eqn} (AD_n CJ + PD_n CJSW)
C_{dbp}	K_{eqp} (AD_p CJ + PD_p CJSW)
C_{gn}	C_{ox} W_n L_n
C_{gp}	C_{ox} W_p L_p
C_L	Σ (C_{gdn}+C_{gdp}+C_{dbn}+C_{dbp}+C_{gn}+C_{gp})

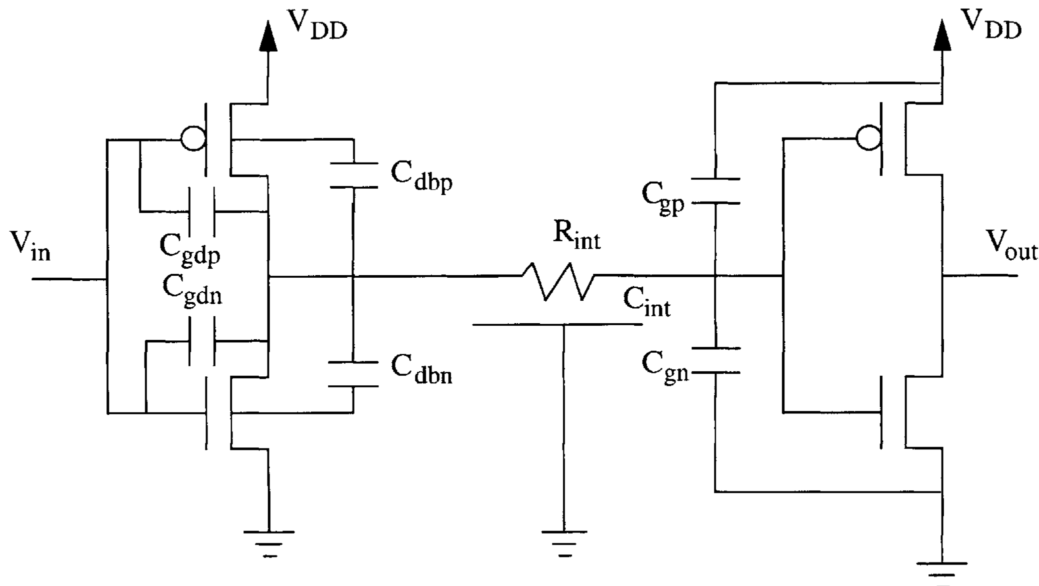


Figure 2.6: The different components of the device load capacitances. The load seen at the driver output is the sum of the gate-to-drain and drain-to-bulk capacitances of the input buffer plus the gate capacitances at the output buffer and the distributed RC interconnect load.

If the interconnect is long enough so that the device load capacitance is much less than the interconnect capacitance ($C_L << C_{int}$), the expression in Eq. 2.6 reduces to

$$T_d = 0.4R_{int}C_{int} + 0.7R_{tr}C_{int}. \tag{2.10}$$

The expression given in Eqs. 2.6 and 2.10 is the Elmore delay, which accounts for the first order moment and is only an approximation. The expressions given here are to provide a qualitative understanding of the effects of the different components involved in computing the signal delay. Spice simulation should be used to include higher order moments when better accuracy is required.

We have seen that the signal delay is a function of the device resistance, load capacitance, and interconnect RC. Variations in the devices affect the driver resistance and load capacitance. For short wires, the devices play a large role in signal delay, and variations in

the devices must be considered. If interconnect resistance dominates over transistor resistance, most of the delay is due to the interconnect RC, and variations in the interconnect geometry can impact the path delay significantly. If this is the case, accurately modeling the interconnect is essential for achieving maximum performance as well as ensuring that specifications are met.

2.3 Clock Skew

Unlike signal delay, the main objective in clock tree design is *matching* the signal paths within the chip. The amount of clock skew depends on the design itself as well as process variations. An important part of design is the degree of asymmetry in the circuit. These asymmetries may be due effects such as differences in the path lengths or load imbalances (see Fig. 2.7). Although it is desirable to make the tree completely symmetric, this may not always be possible in a large chip with several functional units. This is particularly difficult if there are areas of very high density such as SRAMs, and the clock may need to be routed around such blocks. The second cause of clock skew is process variations, both in the interconnect as well as devices. The device variations may impact the output buffers and any intermediate buffers that are used to drive the clock signal.

As we discussed in Chapter 1, the interconnect in an H-tree is tapered so that when a fork is encountered, the linewidth is reduced by a factor of two for impedance matching. We define the number of levels N in a tree as equal to the number of forks encountered in tracing a path from the driver to the output. For example, the H-tree in Fig. 2.7 (b) contains two levels. The number of distinct paths is then equal to 2^N and the number of branches is computed as

$$Branches = \sum_{i=1}^{N} 2^i . \tag{2.11}$$

For example, a two-level H-tree has $2+2^2=6$ total branches or segments.

We compute the signal delay from the driver to output A in this two level H-tree. The signal delay in a tree depends on the resistance and capacitance of all branches in the path as well as the capacitances of all other nodes that contain any of those branches common to the path of interest [2]. This is because those resistances charge (or discharge) all capacitors between the driver and the outputs. Therefore, assuming that the load capacitances are much smaller than the interconnect load, the signal delay from the driver to the output A is given as

$$T_A = 0.7\left(R_{tr}\sum_{i=1}^{6} C_i\right) + 0.4\left(R_0\sum_{i=1}^{6} C_i\right) + 0.4(R_1(C_3 + C_4)) + 0.4(R_3C_3). \qquad (2.12)$$

Since the delay of a given path is dependent on the capacitances of branches not in that path, variations in the interconnect geometry of other branches will also affect the delay. The skew between paths is then computed as the difference in the delays.

2.4 Crosstalk Noise

Crosstalk noise (an induced voltage on a nominally quiescent line) depends on the switching activity of nearby signals. Fig. 2.8 (a) shows the case of a single line switching next to a quiet neighbor. The 2D cross-section is shown in Fig. 2.8 (b) and its equivalent circuit is given in Fig. 2.8 (c). The crosstalk noise depends on the coupling and ground capacitances, the driver and line resistances, aggressor signal rise time, and the supply voltage. Crosstalk noise is generally measured as a percentage of the supply voltage and usually up to 10%-20% noise is acceptable. However, low swing designs have a much lower noise margin and for these cases a much lower crosstalk noise may be required for proper circuit operation. Several publications have derived analytical models for the

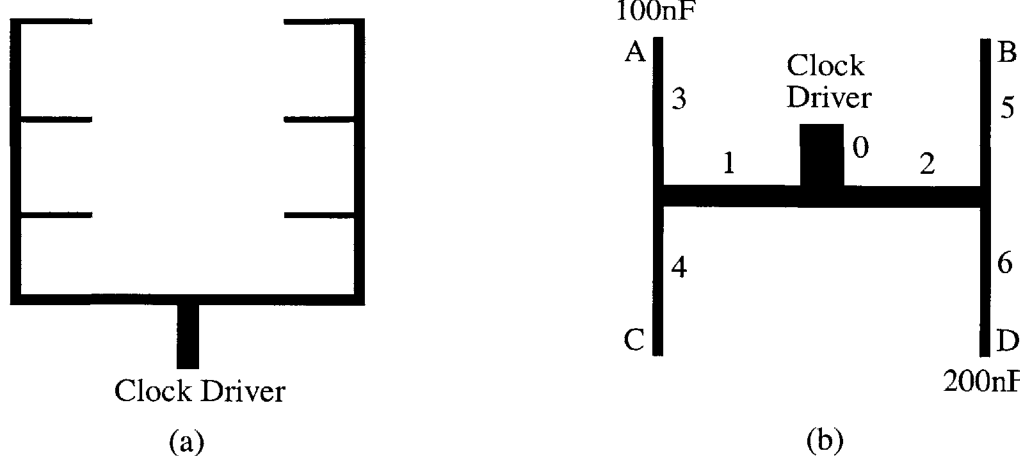


Figure 2.7: The clock distribution network using a spine configuration (a) automatically causes skew due to a difference in the interconnect lengths between the clock driver and the outputs. The H-tree (b) has clock skew because of a difference in the loads, although the distance between the driver and all loads is the same. Both cases are examples of clock skew due to design asymmetry.

crosstalk noise, e.g. [27-30]. The simplest expression for the crosstalk voltage induced on the victim line is based on a ratio of the coupling capacitance C_c to total capacitance (sum of coupling and ground capacitance C_a) and is given by Sakurai [24] as

$$V_{2,max} = \left(\frac{V_{DD}}{2}\right)\left(\frac{C_c}{C_a + C_c}\right). \quad (2.13)$$

This equation assumes that the rise time of the input signal is zero, and does not take into account a line length dependence. For the case of non-zero rise time, modified equations for an RC line are given in [29]. These are more realistic and include a rise time and interconnect length dependence. Crosstalk noise increases for interconnect lengths over 1 mm and begins to saturate around 10 mm. Crosstalk increases with faster rise times and saturates to the expression given in Eq. 2.13 for a step input.

As technology scales, interconnect design must include the effects of crosstalk noise very carefully due to faster switching circuits and longer interconnects. The maximum

crosstalk noise occurs when two adjacent neighbors near a quiescent line are switching. Since the signal states are dynamic, the design constraint must account for this worst-case switching effect. Note that compared to variation, switching activity impacts crosstalk noise more significantly. If one neighbor is switching near a quiescent line, Eq. 2.13 results in a crosstalk noise estimate of $0.25V_{DD}$ compared to $0.5V_{DD}$ if two neighbors are switching simultaneously (doubling of the crosstalk noise), assuming that lateral coupling and overlap capacitances are equivalent. On the other hand, the impact of +20% metal linewidth variation (which increases the lateral coupling capacitance by approximately 20%) results in an increase in crosstalk noise of about 10%. Therefore, if design specifications are tight and noise margin is a concern, interconnect geometry variation may result in an unacceptably high level of crosstalk noise. However, a worst-case *geometry* variation model may not be necessary. The use of variation modeling (especially CD variation) may help reduce this uncertainty (or determine an acceptable inter-wire spacing) and enable a more aggressive interconnect design.

2.5 Inductance Modeling

At high enough clock frequencies, transmission line effects become important and the inductance must also be included for circuit simulation. The defining characteristics of a transmission line are its characteristic impedance and the velocity of propagation [2]. For a lossless transmission line, the characteristic impedance Z_0 of an interconnect is given as

$$Z_0 = \sqrt{\frac{L}{C}} \tag{2.14}$$

where

L = lumped inductance of line,

C = lumped capacitance of line,

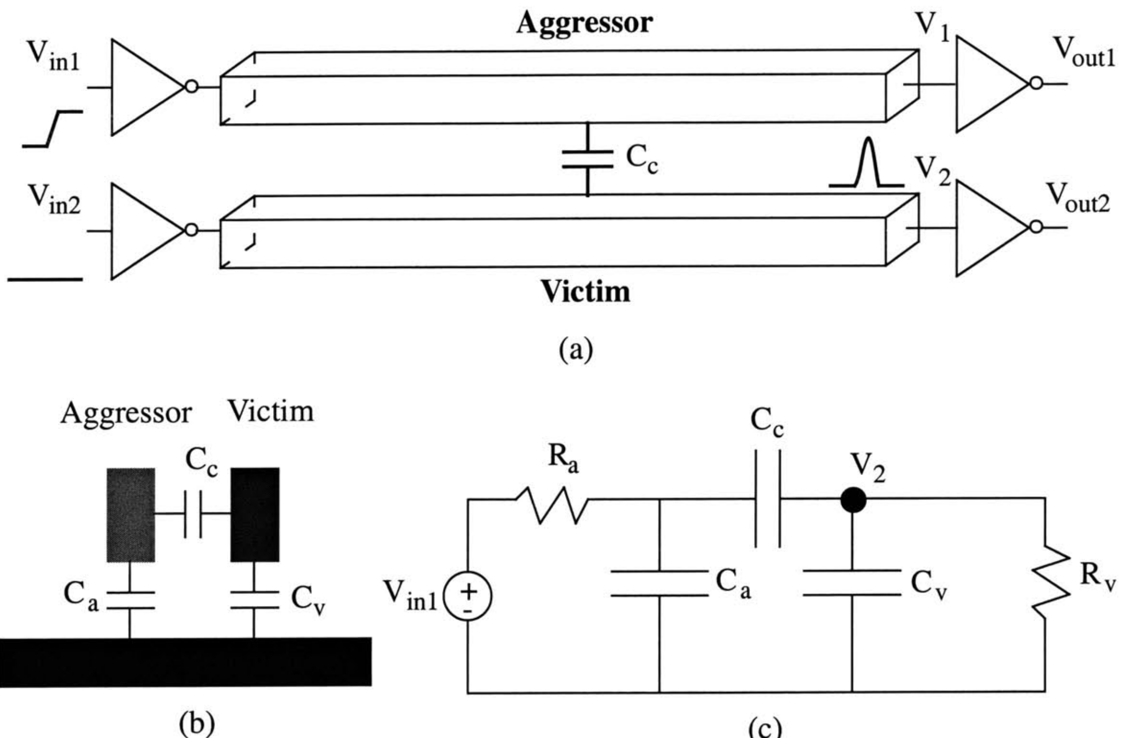


Figure 2.8: An aggressor signal switches next to a quiescent line (a) and results in crosstalk noise at the victim output. A 2D cross-section showing the capacitances of the victim and aggressor signals (b) and the corresponding equivalent circuit used to compute the maximum crosstalk noise $V_{2,\,max}$ (c).

and the velocity of propagation v is calculated as

$$v = \frac{l}{\sqrt{LC}} = \frac{c_0}{\sqrt{\varepsilon\mu}} \tag{2.15}$$

where

l = length of transmission line,

c_0 = speed of light in vacuum,

ε = dielectric constant,

μ = magnetic permeability.

The time of flight delay across the transmission line is given as

$$t_f = \frac{l}{v} = \sqrt{LC}. \quad (2.16)$$

The above equations are for a lossless transmission line. If there are resistive or skin effect losses, these must be included. For on-chip interconnects that behave as transmission lines, resistive effects must be included and an RLC model (instead of LC) must be used. Also, from Eq. 2.5, the skin effect may be important at high frequencies for wide lines. Both of these types of losses result in an attenuated signal.

The inductance becomes important if the signal rise time at the transmission line input is very fast compared to the time of flight delay. Specifically, from [2], transmission line effects must be included if $t_r < 2.5t_f$ and do not need to be included if $t_r > 5t_f$. For $2.5t_f \le t_r \le 5t_f$, transmission line analysis may or may not be needed. The signal rise time depends on the rise time at the driver input and the relative magnitude of the driver source resistance R_{tr}. If the source resistance is much greater than the transmission line characteristic impedance, the rise time is generally slow enough that transmission line effects do not affect the signal delay. If the source resistance is small (i.e. a large driver) compared to the transmission line impedance, the rise time may be small enough that inductance effects need to be included.

2.5.1 Inductance Calculation

The inductance computation is different from the capacitance computation in an important way. The mutual inductance is a function of the inductive loop, which depends on the current path. Since the return path may not be adjacent to the signal wire of interest, longer range interactions need to be considered. Additionally, a signal may have more than one return path where mutual inductance across different conductors needs to be consid-

ered. The inductance may be calculated using numerical solvers such as Raphael or FASTHENRY [31] or approximated using closed form equations [32].

Although inductance calculation is more complex, two facts help make the computation easier. The first is that the self inductance of a wire depends only on its length, width, and height. The second is that the mutual inductance between two wires depends only on the geometry of those wires and is not affected by any other wires. Closed form expressions for self and mutual inductance [32] are shown here for select cases. The self inductance of a wire of length l is given as

$$L(nH) = 2l\left(\ln\left(\frac{2l}{W+T}\right) + 0.5 - k\right) \quad (2.17)$$

and the mutual inductance for two identical wires is given as

$$L(nH) = \frac{\mu_0 l}{2\pi}\left(\ln\left(\frac{2l}{S}\right) - 1 + \frac{S}{l}\right) \quad (2.18)$$

where $k=f(W,T)$ with $0<k<0.0025$, and W, T, and S are in units of cm.

There is an interesting point to note about inductance. Since the mutual inductance depends on the inductive loop, the effects of process variation are not likely to change the mutual inductance significantly between wires that are far away. Therefore, for these cases, an *RLC* line should include the effects of geometry variation on R and C, but does not need to be concerned as much with the effects of variation on L.

In this thesis we focus on the impact of variation in *RC* interconnect, but our methodology can be extended to *RLC* lines when necessary. The next chapter provides an overview of systematic variation modeling.

Chapter 3

Process Variation Models

In this chapter, we describe layout dependent interconnect variation models based on spatial patterns and the type of interconnect process used in a given technology. We will see that the process has a large impact on the interconnect variation. Major sources of interconnect variation are from chemical mechanical polishing (CMP) as well as linewidth variations during patterning. Other types of potential variation sources, such as film thickness deposition and material properties (e.g. metal resistivity and insulator dielectric constant) are usually well controlled. Therefore, our focus is on the impact of metal or ILD thickness variation resulting from CMP and linewidth variation in the metal or polysilicon lines.

First, we discuss interconnect variation models for CMP planarization. From Chapter 1, we know that the ILD thickness variation in an oxide CMP process has a pattern density dependence. This chapter discusses the ILD thickness variation model in more detail. We also look at interconnect variation in a damascene CMP process. Here, the oxide is patterned rather than the metal. This is followed by metal deposition and metal CMP. Since the metal is polished instead of the oxide, the CMP variation is in the metal thickness (and not the oxide). A damascene CMP process is used with copper (Cu) interconnect since copper is much more difficult to pattern and etch than aluminum (Al) interconnect. Since copper has a lower resistivity than aluminum, the trend is moving toward copper wires starting around the current day (180 nm) generation. Characterizing and modeling the copper CMP variation behavior is therefore a very active area of research [33-36].

We also consider linewidth variation in metal lines (interconnect) as well as polysilicon (devices). The metal linewidth and polysilicon device channel length are often referred to as the critical dimension (CD). The CD variation is a strong function of the layout patterns [37]. The CD variation in metal lines affects intra-layer coupling and variation in poly CD affects device delay. Several works have studied the behavior of CD variation (e.g. [37-38]), but relatively few models exist because of the complex nature of linewidth variation. We review some of these previous works and discuss the main factors that cause CD variation.

After a description of ILD CMP modeling, we proceed to look at metal CMP modeling. We then discuss the systematic sources of metal and poly CD variation. Finally, we provide an example to show the importance of including systematic models in circuit simulation by comparing the effects of pattern density on interconnect delay using two different layout patterns.

3.1 An ILD CMP Model

The within-die ILD thickness variation is a function of the CMP process and the interconnect layout geometry and its surroundings. The ILD thickness variation model is based on Preston's equation, which relates the removal rate on a blanket wafer to the pressure velocity product. We discuss the basic derivation of the ILD CMP model proposed by Stine et al. [20]. Consider the structure shown in Fig. 3.1. The blanket (planar region) removal rate is given as

$$RR = \frac{dz}{dt} = -\kappa P v \tag{3.1}$$

where P is the pressure, v is velocity, and κ is a proportionality constant. Differences in pattern density result in varying amounts of post CMP ILD thickness across the chip (see

Fig. 1.11) since sparse regions polish faster than dense regions. The (effective) pattern density is calculated as the amount of metal in a given area divided by the total area in that region (see Fig. 3.2). The "given area" refers to the planarization length or interaction distance which is a function of the process, CMP tool, and consumables. The pressure can be represented as F/A, where F is the down force exerted on the wafer and A is the area of the oxide contacted by the pad. With id defined to be the interaction distance and $\rho(x, y, z)$ being the effective pattern density,

$$\frac{dz}{dt} = \frac{-\kappa F v}{(id)^2 \rho(x, y, z)}. \tag{3.2}$$

This gives

$$\frac{dz}{dt} = \frac{-K}{\rho(x, y, z)} \tag{3.3}$$

where

$$K = \frac{\kappa F v}{(id)^2} \tag{3.4}$$

is the blanket oxide removal rate. The removal rate for a given location at a specified time depends on whether or not there are oxide features remaining above the metal, giving rise to two polishing regimes. The pattern density for each of the regimes is calculated as

$$\rho(x, y, z) = \begin{cases} \rho_0(x, y), & z > z_0 - z_1 \\ 1, & z < z_0 - z_1. \end{cases} \tag{3.5}$$

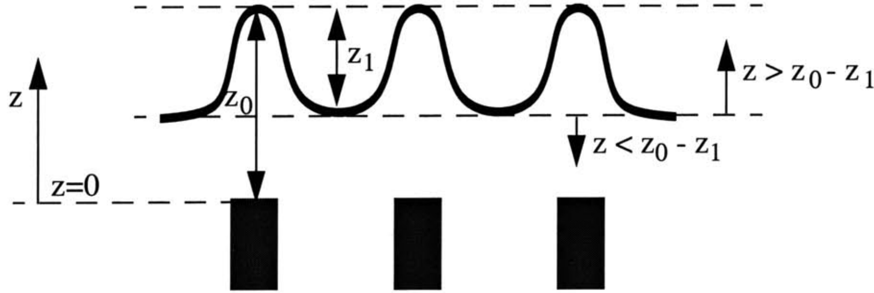


Figure 3.1: The oxide polish rate depends on the underlying metal pattern density until the oxide features ("up areas") have been removed. In this regime, there is negligible polishing of the "down areas". In the second regime (when all up areas are removed), the oxide polishes at the blanket removal rate.

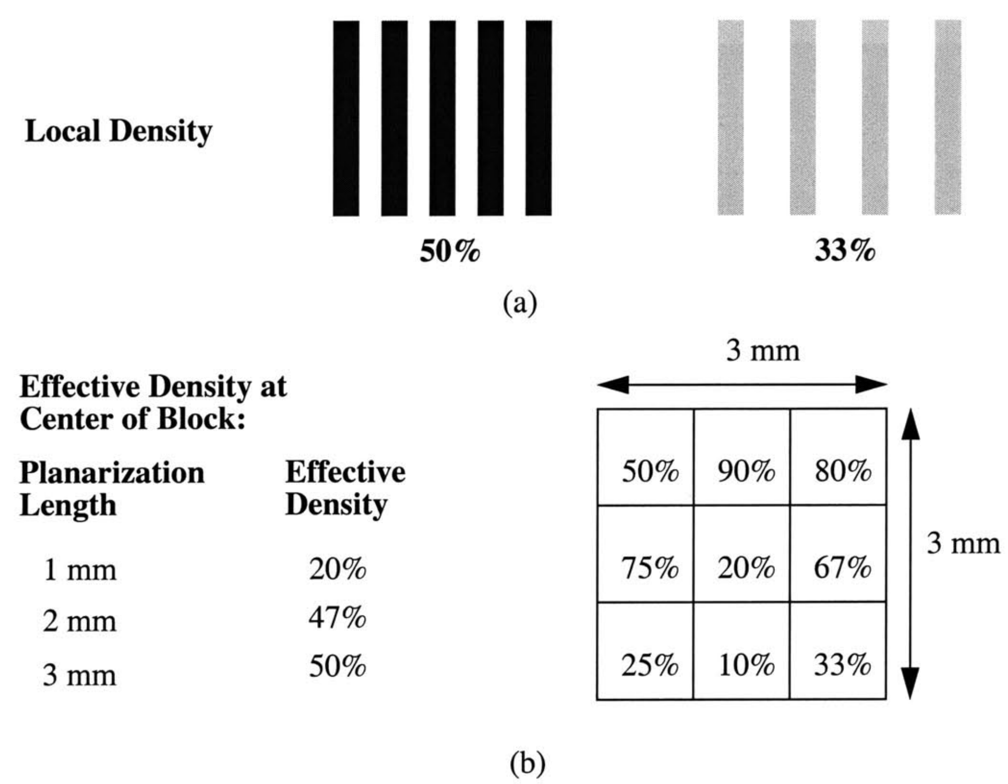


Figure 3.2: The local density is taken as the ratio of linewidth to pitch (a). The effective pattern density depends on the planarization length (b). A larger interaction distance results in more averaging across the chip.

For a fixed polish time, the post CMP ILD thickness can be calculated across the entire chip. If the target polish time is such that all features have cleared, the ILD thickness above that metal layer can be calculated across the entire chip as:

$$ILD_{final} = ILD_{nom} + (\rho - \rho_{nom})z_0 \tag{3.6}$$

where

ILD_{final} = ILD thickness after planarization at density ρ,

ILD_{nom} = ILD thickness after planarization at density ρ_{nom},

ρ = effective pattern density calculated using planarization length,

ρ_{nom} = effective pattern density at target dielectric thickness ILD_{nom},

z_0 = as-deposited step height.

Thus, if the effective density range across the chip is $\Delta\rho = \rho_{max} - \rho_{min}$, then the ILD thickness is $\Delta\rho \cdot z_0$. A 50% density range with a 0.6 μm step height can give rise to a substantial 0.3 μm ILD thickness variation.

3.2 A Copper CMP Model

A damascene CMP process is typically used with copper interconnect (see Fig. 3.3). Unlike an aluminum interconnect process where the metal is patterned and the oxide is polished, the oxide is patterned and the metal is polished in a damascene process. This results in systematic variation in the metal thickness, while no systematic effect on ILD thickness variation is observed. The metal thickness loss is due to two effects known as metal dishing and oxide erosion, shown in Fig. 3.4. There is a high amount of dishing for wide lines, and erosion increases with increasing metal pattern density. Erosion generally dominates over dishing for fine pitch lines, especially at high density.

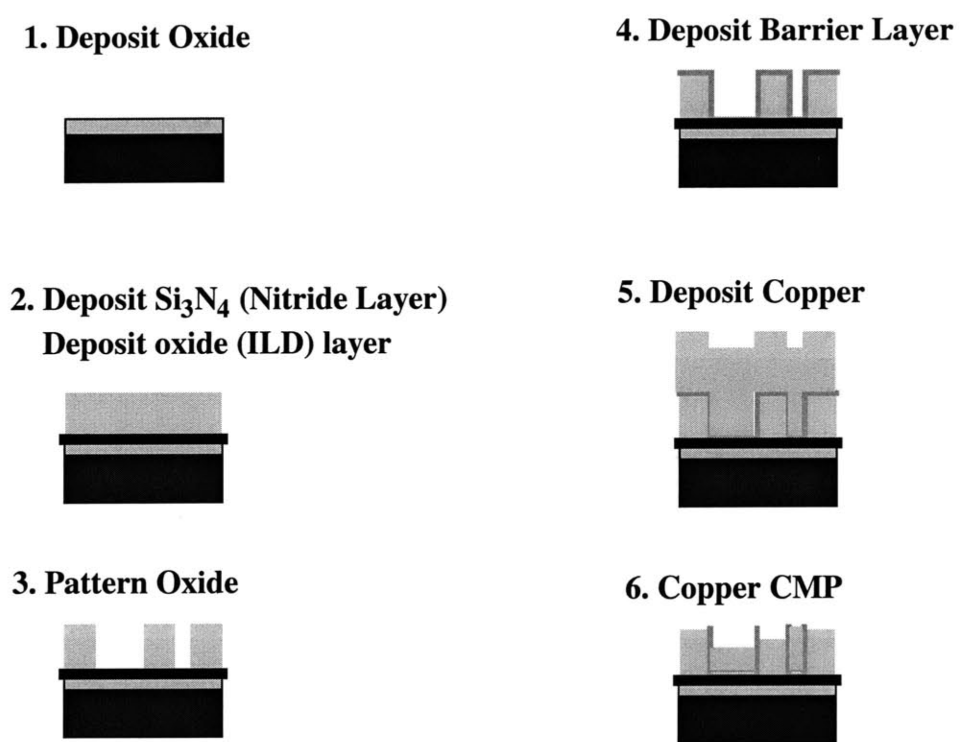


Figure 3.3: A damascene CMP process used with copper interconnect. The oxide is first patterned, followed by metal stack deposition. The metal is polished until all "up areas" have been removed. To guarantee that there are no inadvertent shorts between adjacent wires, overpolishing is done resulting in metal thickness loss.

The copper CMP metal thickness variation depends on the layout pattern and several process factors. It is much more complex than the ILD thickness variation model and there is ongoing research in the area of copper CMP modeling to explain the pattern and process effects on metal thickness variation. A good description of a mathematical model is given in [34], where a time dependent model describing the CMP process is provided for three different polishing stages. In other work [35-36], substantial data has been collected using test masks that contain various combinations of layout patterns.

Although variations in process conditions impact the post CMP metal thickness variation, our goal in this thesis is to develop methods to understand the effect of important process variations on circuit performance. For our purposes here, we consider a model that is

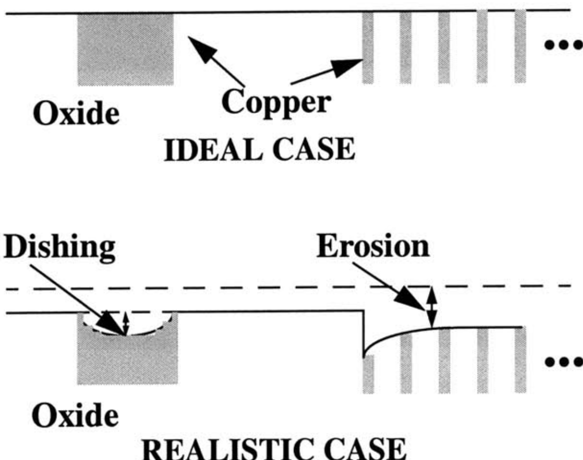


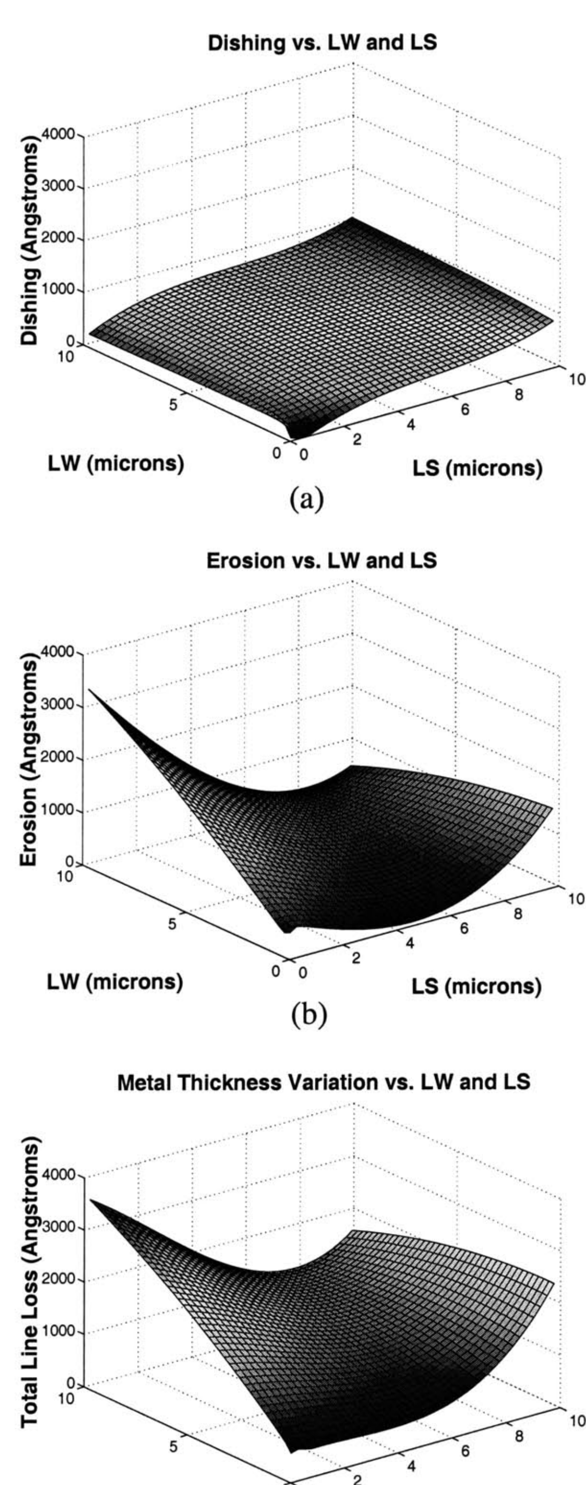
Figure 3.4: Overpolish in copper CMP processes results in metal dishing and oxide erosion. Dishing is present in wide lines and erosion dominates for fine pitch lines at high density, both causing metal thickness loss.

based on data taken from an MIT/SEMATECH mask [39]. The copper CMP metal thickness variation is modeled as a function of both linewidth and linespace. The data used in the model is for a minimal (~10%) overpolish with a polish time of 107 seconds using a rotary polishing tool. Fig. 3.5 shows the dishing, erosion, and total metal thickness loss.

In terms of the effect on electrical parameters, an ILD CMP process only causes variation in the inter-layer dielectric and therefore only affects the inter-wire capacitances. With metal CMP, metal thickness loss results in an increase in the resistance. It also decreases the lateral intra-layer coupling among adjacent wires. To some extent, the increase in resistance is offset by a decrease in the capacitance. How much the metal CMP variation increases or decreases the interconnect delay depends on the dimensions of the interconnect structure, and a pattern dependent model is necessary to determine the metal thickness loss and then compute the effect on the delay.

3.3 Critical Dimension Variation

Metal and polysilicon critical dimension (CD) variation is another important issue. The critical dimension refers to linewidth variation in the interconnect and polysilicon



(c)

Figure 3.5: Metal dishing (a), oxide erosion (b), and total metal thickness loss (c) as a function of linewidth and linespace. Erosion is the dominant component for the range of *LW* and *LS* combinations shown here.

gate length variation in the devices. Variations in the interconnect length and the device width have a much smaller impact. Typically, fine pitch wires have the greatest impact on the interconnect delay. These are representative of most interconnect in the local and intermediate wiring tiers. Although local wires are not used to route over long distances across the chip, metal linewidth variation may still be a concern. Also, with technology scaling, the global wiring pitch is shrinking, making the wires more susceptible to metal CD variation. The main concern with interconnect CD variation is crosstalk noise and signal delay. Clock wires are typically much wider than minimum pitch and are affected less in terms of skew. They are also well shielded and not as likely to be impacted by neighboring signals. Poly CD variation, however, can have a big impact on the buffer delay and is important for all three design metrics considered here (signal delay, clock skew, and crosstalk noise). Since matching the signal arrival times is the essence of a good clock design, poly CD variation must be considered at the clock outputs and any intermediate buffers.

The CD variation in both metal and poly lines is due to the same physical characteristics and is caused by mask, lithography, and etch effects [37-38, 40-42]. The mask errors are due to patterning the reticle. The lithography effects may be due to lens aberration or stepper leveling and focusing errors, in addition to well known optical proximity effects. The etch effects are generally due to differences in loading or aspect ratio (feature size) dependencies. The CD variation can be relatively large, and test measurements have shown that it can be as high as 15-20% of the minimum linewidth.

Although CD variation can be significant, it can be difficult to model since it depends on many different factors. Additionally, the model needs to be calibrated as a function of the process parameters. Many works have attempted to model the CD variation through analytical models, e.g. [40-41]. One approach relies on physical models to capture optical proximity using aerial imaging simulation. In other studies, test masks with different pat-

terns have been used to look at the trends for CD variation. In [37], Liebmann et al. look at array and pattern density effects, both of which are found to be important. The paper concludes that there are both short and long range effects. Only adjacent neighbor lines need to be considered to account for the short range effect. The pattern density must be considered to account for the long range interactions.

While all the effects of CD variation may not be due to systematic effects, some of this systematic variation can be modeled. Optical proximity correction (OPC) techniques generally incorporate physical or empirical models to correct some of the variation. This is done by applying a positive or negative bias to patterns in the mask layers [42].

In our simulation studies of the impact of CD variation in future chapters, we do not use one specific model. Our goal is to understand the relative importance of including systematic variation models for the metal or poly CD compared to other effects such as CMP models. Therefore, we assume that a fraction of the CD variation can be corrected and the rest of it is random. Note that OPC effectively results in a tightening of the distribution. We assume that this is the case and compute the performance gain that is obtained by tighter tolerance design vs. that of the worst-case tolerance. We then compare these gains with a tighter tolerance for the metal and poly CD vs. gains resulting from systematic modeling of the metal or ILD CMP.

3.4 Example: Systematic vs. Worst-Case Variation Modeling

Although variation can negatively impact circuit performance, knowing how much variation is there and how it effects each circuit or critical path in a high performance design can be extremely useful. Not only can modifications be made to alter the design, but the design margin uncertainty can be reduced with the use of systematic variation models in circuit simulation.

As an example, the delay of 10 mm long interconnect is simulated. The interconnect parameters are taken from the SIA roadmap for a 0.25 μm technology, and a Level 3 Spice model is used. The circuit consists of a minimum sized buffer at the input and output, with the 10 mm wire in between. The simulations are performed for two different cases. In the first case, a fixed minimum value of ILD thickness (1.18 μm) is used. The minimum (worst-case) ILD thickness for the first case is computed by sampling the ILD thickness profile with a fine grid spacing. The reasoning is that the minimum and maximum variation limits may be obtained by sampling the ILD thickness in the laboratory (e.g. through profilometry measurements) for establishing design rules in a similar way. These limits are given in Table 3.1. In the second case, an effective pattern density profile (based on a 4 mm planarization length) is taken for a global wiring layer in a microprocessor fabricated in 0.25 μm technology (see Fig. 3.6 (a)). The ILD thickness is computed using the pattern density model described earlier in this chapter, assuming a target ILD thickness polish time such that the nominal ILD thickness (1.78 μm) corresponds to the case of 50% underlying metal pattern density (see Fig. 3.6 (b)).

The interconnect circuits (100 of them) are randomly placed above the given pattern density profile. As the interconnect travels across the die from the input to the output buffer, variations in the ILD thickness result in an increase or decrease in the capacitance along different sections of the wire. The interconnect is divided into sections of 100 μm in length to model these variations as a function of spatial location as well as to account for the distributed effect of the wire. The resistance is computed directly from Eq. 2.4 and the capacitance is pre-computed using a 2D solver (Raphael) for different pattern densities ranging from 0-100%. Fig. 3.6 (c) shows the Spice simulation results and Fig. 3.6 (d) shows the distribution. The results show that the actual worst-case delay (predicted by the spatial ILD thickness variation model) is 8% less than that computed with the worst-case

ILD thickness since the length of the wire is relatively long and the effects of variation tend to cancel out.

Although pattern dependent modeling results in a moderate improvement for the case study presented above, its effectiveness over the worst-case approach largely depends on the circuit layout. We consider for our next example the pattern density profile in an ASIC. This case study differs from the microprocessor example because of the difference in the pattern density profile as well as the interconnect technology. This ASIC is fabricated using only a three level metal process. In the lower metal layers, the nominal ILD thickness is thinner than the upper layers, so the effect of variation may be more pronounced. Table 3.2 and Fig. 3.7 show the specifics of this example. We find that for the ASIC, using a pattern dependent model results in a significant improvement (26% gain) over the conventional worst-case approach.

These examples show that separating the total variation into systematic and random components reduces the overall uncertainty and allows for greater flexibility in design. Additionally, other sources such as metal thickness and linewidth variation in the interconnect as well as device gate length variation may have systematic components. With increasingly demanding design targets, circuit design in the future must account for realistic within-die variation in both interconnect and devices. In the next chapter, we describe such a methodology for automated variation impact assessment.

Table 3.1: Microprocessor ILD Thickness

Parameter	Value (μm)
Mean ILD Thickness	1.78
Minimum ILD Thickness	1.18
Maximum ILD Thickness	2.38

Microprocessor pattern density data courtesy of Hewlett Packard

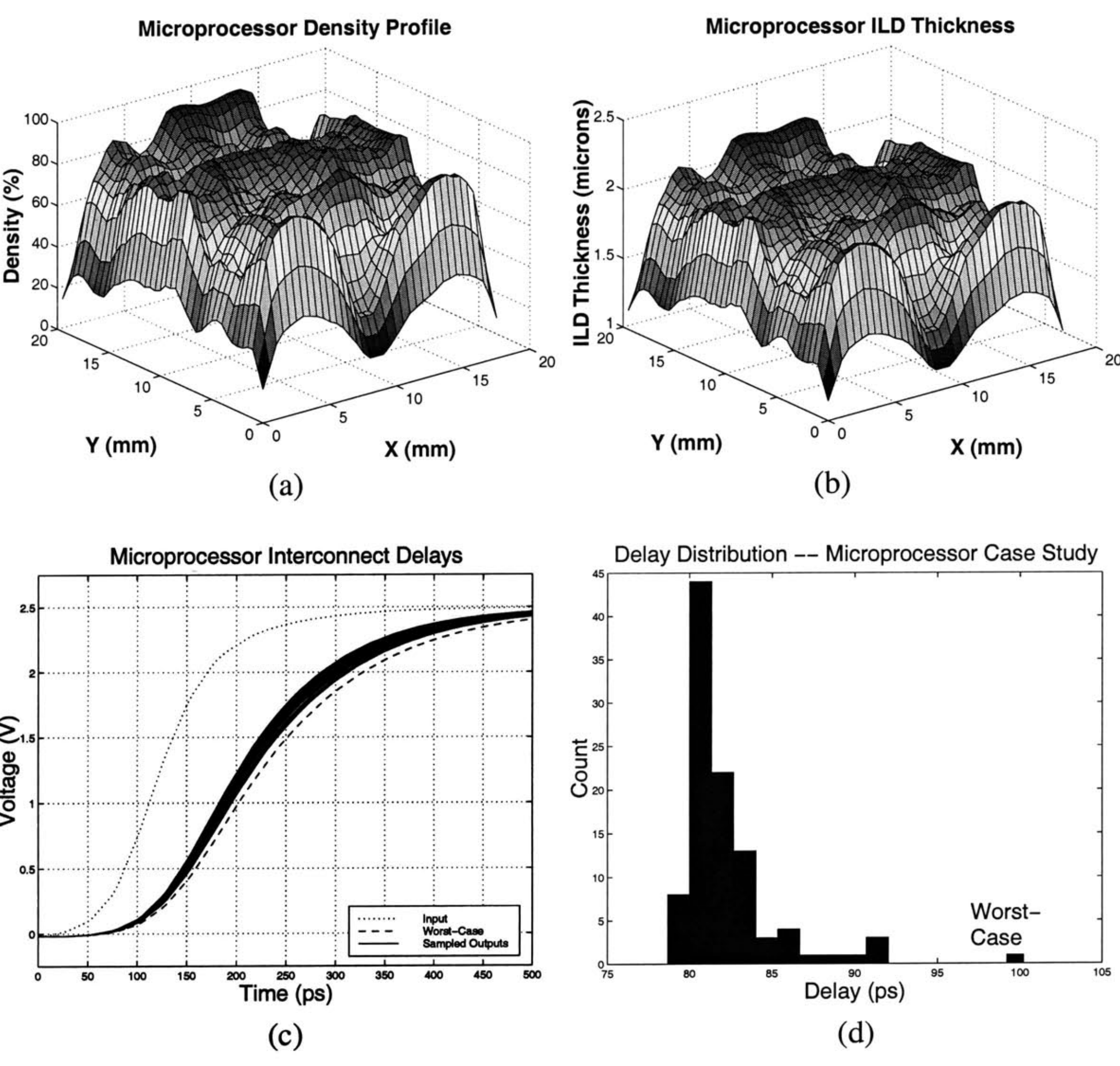


Figure 3.6: The pattern density (a) and ILD thickness variation (b) between the top two metal layers in a microprocessor using an interaction distance of 4 mm. The simulated delays of 100 interconnect lines 10 mm in length placed randomly across the chip (c) and a histogram showing the distribution (d). The worst-case ILD thickness value overestimates the actual worst-case by 8%.

Table 3.2: ASIC ILD Thickness

Parameter	Value (μm)
Mean ILD Thickness	0.82
Minimum ILD Thickness	0.40
Maximum ILD Thickness	1.24

ASIC pattern density data courtesy of Hewlett Packard

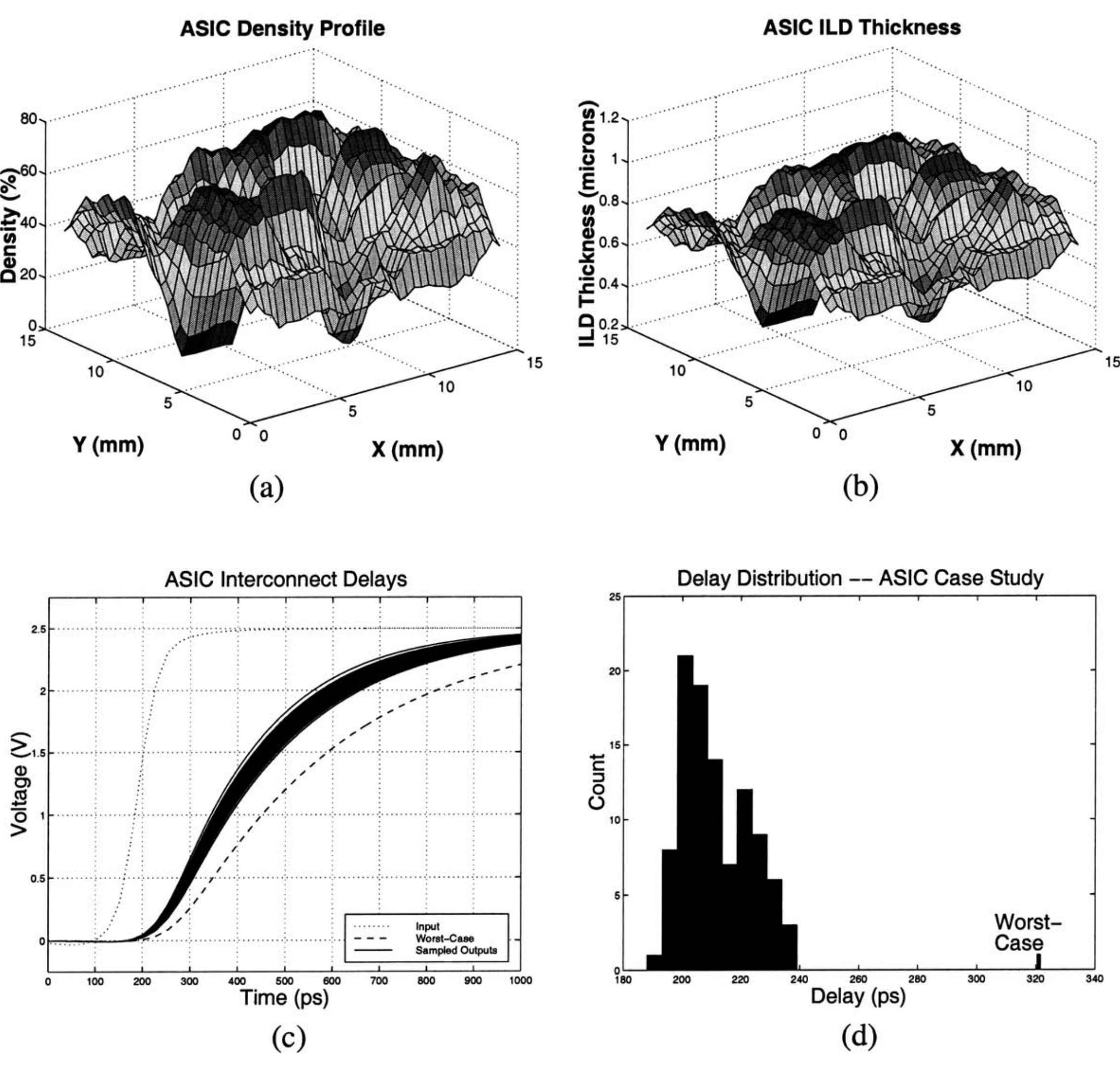


Figure 3.7: The pattern density (a) and ILD thickness variation (b) between the top two metal layers in an ASIC using an interaction distance of 4 mm. The simulated delays of 100 interconnect lines 10 mm in length placed randomly across the chip (c) and a histogram showing the distribution (d). The worst-case ILD thickness value overestimates the actual worst-case by 26%.

Chapter 4

Systematic Variation Analysis Methodology

This chapter describes a methodology for assessing the impact of systematic interconnect and device variation on circuit performance. We have seen in the previous chapters that layout and process effects are important in determining how much variation exists for a given device or interconnect structure. While most circuit (interconnect and device) extraction tools allow a detailed and accurate extraction of circuit parameters to be used for circuit simulation, systematic variational analysis is usually not an option. In particular, most existing parasitic RC extraction tools do not account for within-die interconnect parameter variation. Typically, interconnect technology parameters are fixed by parasitic extractors and a single fixed value is used for the ILD or metal thickness of each layer. While flexibility is given for design parameters (horizontal direction), process parameters (vertical direction) are determined by the technology and assumed to be fixed for a given mask layer. Therefore, a methodology is needed to make use of the knowledge about pattern dependent variation and implement it for circuit simulation.

We start with a review of some earlier works that have looked at systematic variation impact. We then present our new methodology and describe its implementation within a CAD framework for automated variational analysis. Finally, we show how our technique may be used for variation impact *assessment* as well as variation *reduction*.

4.1 Variation Analysis Simulation Methods

Most variation analysis techniques generally consist of worst-case skew corner or Monte-Carlo simulations. Although variants of these techniques have been presented in previous works [7-12, 43-45], these approaches rely on a purely statistical analysis. A cir-

cuit simulation methodology utilizing a Monte-Carlo approach is shown in Fig. 4.1. Variation statistics are input into a circuit simulation tool and random sampling is used to obtain an output distribution. Although statistical analysis techniques have been popular, few studies have dealt with systematic variation effects. In [40], a study is done to assess the effects of poly CD variation on an SRAM using aerial imaging. In [46] pattern dependent copper CMP effects are explored on a clock tree, but the study is limited to metal dishing.

One of the few early works where a methodology is proposed to study the impact of pattern dependent interconnect variation is [47]. The concept of the "net halo" is employed. The interconnect of interest and its surroundings a specified distance away (called the net halo) is extracted. The halo is selected to capture most of the electromagnetic coupling, mainly used for calculating the coupling capacitance. The process variation is modeled separately using a different length scale (e.g. planarization length for pattern density calculation). Using this information, detailed 3D capacitance simulations are performed. While this technique is extremely accurate, it requires a full capacitance simulation with each new process condition. A complete capacitance re-extraction becomes prohibitively expensive for large circuits or to analyze the impact of different variation sources (e.g. LW, ILD thickness) independently.

In this thesis, we present a new method to study the systematic variation impact on circuits [48, 49]. Our technique overcomes the limitations of [47] and does not require a re-extraction of net parameters to model the variation even as process conditions change or new variation sources are considered. We implement this technique within a CAD framework to automate the circuit performance simulation and assess the impact of different types of variations on high performance microprocessor designs. The rest of this chapter describes our new methodology.

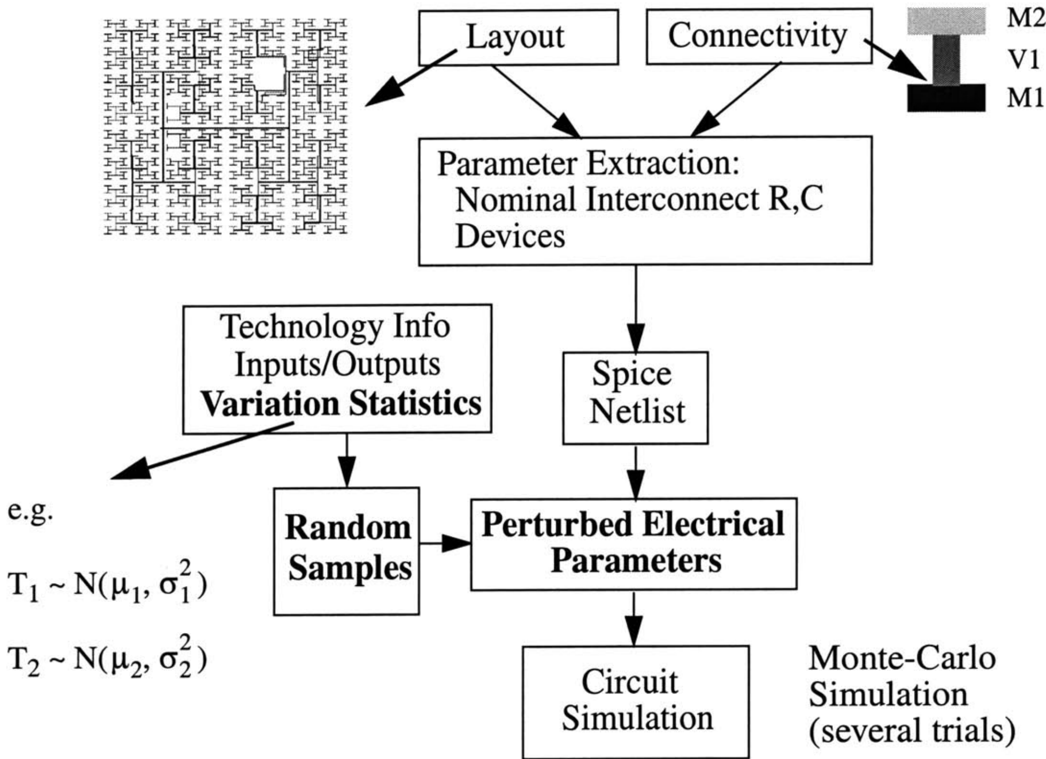


Figure 4.1: The conventional approach to modeling the effects of variation using statistical circuit analysis. Variation statistics are input for each parameter and Monte-Carlo simulations are performed to obtain a delay distribution.

4.2 Net Extraction

In order to study variation impact, an automated methodology is needed to efficiently study the effects on circuit performance. A different approach is needed if there are a small number of regular, symmetric structures than if thousands of irregularly shaped wires need to be analyzed. With a small number of structures, the technique is straightforward, and extremely accurate 3D capacitance simulations may be performed. For the case of an entire chip such as a microprocessor, the technique must be integrated within a CAD tool framework to efficiently analyze thousands of nets. Some important considerations

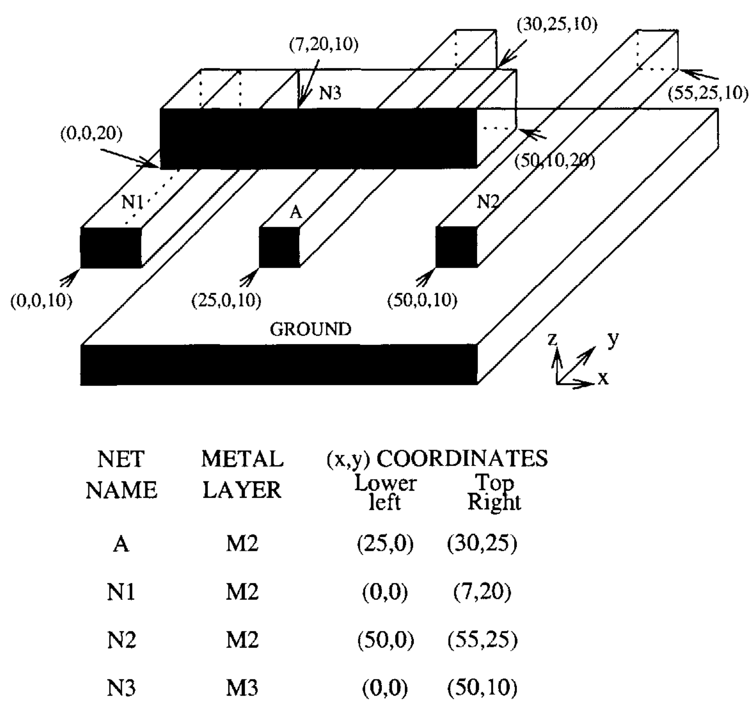
include compatibility with existing net and capacitance extraction tools and efficiency in capacitance re-calculation with changing process conditions.

In order to study the effects of parameter variation, capacitance extraction tools must be flexible enough to handle as *variables* many of the technology parameters that are typically fixed by most conventional extractors. Ideally, a capacitance extractor would calculate variation sensitivity in addition to the nominal capacitance for any parameter. These include parameters such as the ILD thickness or metal thickness, which are generally set to a single fixed value for each metal layer. Most tools do not explicitly report sensitivity information, nor do they accommodate die position dependent technology parameters such as layer thicknesses. Modification of commercial extraction tools is slow or difficult. Therefore, without modifying the extraction tool internally, an interface is required between the extractor and external process variation models.

To make use of pattern dependent process variation models, a key feature is necessary in the new proposed methodology -- interconnect *geometry* and *coordinates* must be known for the net or device of interest and its surrounding neighbors (see Fig. 4.2). This information must be output by the extractor to effectively model pattern dependent effects. First, the position on the chip is used to determine the variation in the geometric structure of the interconnect or device parameter based on process variation models. Second, the geometry information is used to calculate the resistance and capacitance variation.

An important consideration in using the geometry and coordinate information effectively is that the interconnect segments must be small enough that a specified *location* can be assigned to the segment to model the variation at that location. Typically, nets are segmented into small sections automatically by the extractor due to changing neighbor environments along different parts of the net. An analysis of the metal 5 layer of the clock net from an IBM microprocessor [50] shows that the average length of a net segment is 18.8

μm. The interaction distance used to model the effects of CMP is on the order of several mm (an oxide CMP process) or hundreds of microns (a copper CMP process), so assigning a specific location to a net segment is an effective method of modeling the effects of systematic spatial variation.



NET NAME	METAL LAYER	(x,y) COORDINATES Lower left	Top Right
A	M2	(25,0)	(30,25)
N1	M2	(0,0)	(7,20)
N2	M2	(50,0)	(55,25)
N3	M3	(0,0)	(50,10)

Figure 4.2: In addition to the net capacitance or resistance, information about the geometry and coordinates of each net and its surrounding neighbors is also stored during the extraction. This information is then used to predict the spatial variation (e.g. ILD or metal thickness variation) and compute the change in electrical parameters.

Interconnect capacitance extraction involves segmenting the wire into several small sections to model the structure as a distributed RC network. In order to reduce simulation times, a variety of approaches may be used to study the effects of variation. One approach is to first calculate the *nominal* interconnect capacitances accurately using a combination 2D/3D extraction. If enough information about neighbor surroundings is stored in the output file during the nominal capacitance extraction, the change in capacitance (delta capac-

itance, ΔC) due to variation may be approximated through models or capacitance formulas [23-25]. The delta capacitance and resistance are then computed as

$$\Delta C = \hat{C}(W, S, T, H) - \hat{C}(W + \Delta W, S + \Delta S, T + \Delta T, H + \Delta H) \quad (4.1)$$

and

$$\Delta R = \hat{R}(W, T) - \hat{R}(W + \Delta W, T + \Delta T). \quad (4.2)$$

Here, $\hat{C}(W, S, T, H)$ and $\hat{R}(W, T)$ are computed using formulas to approximate the *extracted* capacitance ($C(W, S, T, H)$) and resistance ($R(W, T)$). The delta capacitance and resistance are then added to or subtracted from the nominal extracted values. If the effect of different process conditions needs to be simulated, new values of the interconnect geometry can be computed using the coordinates of the net of interest and its surroundings and applying an updated process variation model. No new RC extraction or net annotation is required. The accuracy of our approach depends on how much information about the surrounding nets is stored and the accuracy of the closed form capacitance models that are used. The amount of information stored about neighboring nets can be varied and is specified by providing the halo distance. If some of the critical paths require a higher accuracy, a larger distance may be used to capture coupling effects further away. For most nets, the nearest neighbors are sufficient for capacitance calculations.

4.3 New Methodology for Variation Assessment

A flowchart for our proposed methodology to study the effects of systematic pattern dependent variation on circuit performance is shown in Fig. 4.3. Our technique enables us to study the impact of any type of spatial variation, encompassing both systematic and random components. Although our focus is on interconnect, spatial device variation models

can also be included. The methodology is implemented within a CAD tool framework (see Fig. 4.4) for automated circuit performance analysis and is reasonably compatible with most existing circuit extraction and simulation tools. We now describe each of the blocks in our flowchart.

(1) Layout

The layout is taken as input for the extraction tool. The layout format is a GDSII or CIF file that is streamed out from a layout editor such as Virtuoso [51] by Cadence.

(2) Layer Connectivity

The connectivity information refers to the order of connection of the different mask layers. This includes the definition of poly, metal, and via mask layers and their connectivity. This information needs to be specified since the interconnect extraction tool must know how each of the layers are connected.

(3) Parameter Extraction

The nominal interconnect parasitics (resistances and capacitances) and devices are extracted using the layout and connectivity information. A netlist is created along with the geometry and coordinate locations. A variety of extraction tools may be used here such as the Cadence Layout Parasitic Extraction (LPE) tool Dracula [52], 3D extraction tools such as Raphael, or other tools. The main requirement for our methodology is that geometry and coordinate locations of the interconnect segments or devices be output during the extraction, in addition to the node names. Our implementation is compatible with two different tools: a proprietary IBM 3D capacitance extraction (3DX) tool [53] as well as for the commercially available Cadence Dracula extractor.

(4) Variation Analysis Tool

This tool is the basic implementation of our methodology. It consists of perl scripts [54] and Matlab functions [55] written to parse the various files and modify the extracted

nets and devices. The inputs to this tool are the extracted nominal interconnect and device parameters, technology information, pattern dependent variation models, signal input/output nodes, and pre-computed pattern densities of all metal layers. The technology information refers to nominal values of the metal and ILD thickness for all layers.

The variation analysis tool perturbs the extracted interconnect resistances, capacitances, and any pattern dependent device parameters using spatial information. The interconnect (and device) geometry variation is calculated based on the pre-computed pattern density, linewidth, and linespace to the nearest neighbors using pattern dependent variation models. Electrical parameter variations are computed next, and closed form expressions are used to calculate the interconnect capacitance variation. Critical nets at or near the target specifications may be fine tuned subsequently using a full 3D capacitance solver. Although we specifically deal with pattern dependent intra-die variation, other systematic variation components can also be incorporated into our methodology. Random variations can then be considered separately using conventional statistical analysis techniques.

(5) Pattern Densities

The effective pattern densities are pre-computed for each metal level, given a specified interaction distance. The local pattern densities are first extracted for a square grid over a small range (e.g. 100 μm) using a density extractor. Our implementation uses the SiCat tool [56] from PDF Solutions to perform these extractions. To compute the effective pattern density, Matlab scripts are written to perform moving average or filtered effective density computations, given a variable planarization length. A more detailed description of the pattern density calculation technique is provided in Section 4.3.1.

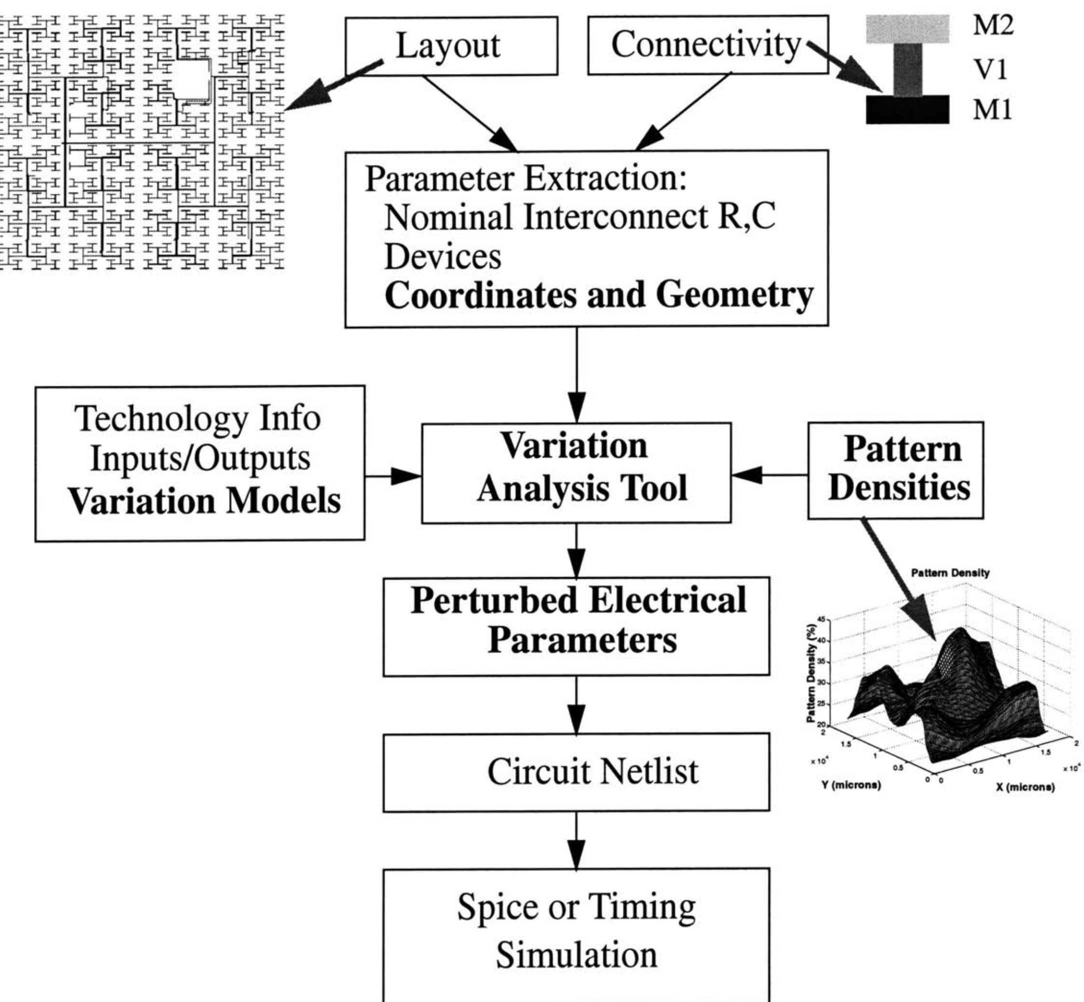


Figure 4.3: The methodology used to simulate the effects of pattern dependent interconnect and device variation on circuit performance. Elements in **bold** are used to model the effects of variation.

(6) Perturbed Electrical Parameters

A modified netlist containing perturbed electrical (and device) parameters is formed as the output of the variation analysis tool. An Hspice [57] simulation can then be run to compute the circuit performance metric of interest. This technique is compatible with other circuit simulators as well, including timing analyzers and model order reduction approaches to reduce circuit simulation time [58]. For the case of the IBM capacitance extractor, the netlist is compatible with a timing analyzer.

Procedure for systematic variation impact assessment:

1. Read in layout file (CIF, GDS).

2. Specify layout connectivity order.
   connect m1 m2 by v1
   connect m2 m3 by v2
   etc.

M2
V1
M1

3. For all nets of interest, extract nominal values of:
   (a) Interconnect resistance
   (b) Overlap capacitance
   (c) Lateral capacitance
   (d) MOSFET devices

4. Sample extraction output:
   R1 N1 N2 20.0 **L=10.0 W=0.5 X=2750 Y=3000**
   R2 N3 N8 10.0 **L=20.0 W=2.0 X=1005 Y=220**
   ....
   C1 N7 GND 1.2e-16
   C2 N3 GND 5.0e-17
   ....
   C2000 N4 GND 3.5e-17
   C2001 N25 GND 4.0e-16
   ....
   MN1 N10 N11 N9 N9 NMOS L=.50u W=4u **X=9040 Y=3056**
   MP1 N10 N11 N6 N6 PMOS L=.50u W=8u **X=9040 Y=3070**
   ....

5. Run variation analysis tool.
   Inputs:
   (a) Technology information:
      Device models
      Interconnect technology parameters
         e.g. Nominal metal and ILD thickness for all layers
   (b) Net input/output nodes
   (c) Nominal values of electrical parameters from step 4
   (c) Variation models
      ILD CMP, Cu CMP, etc.
   (d) Pre-computed pattern densities for all mask layers
   Output: Spice netlist with **perturbed electrical parameters**

e.g.
M5
M5 Pattern Density
M6
M6 Pattern Density
Y (microns)
X (microns)

6. Perform Spice simulation and analyze delays.

Figure 4.4: The procedure used to implement systematic variation models in a CAD tool and automate performance impact assessment.

4.3.1 Pattern Density Calculation

To model the effects of CMP variation, the effective pattern density must be known for each interconnect segment. The local density is first computed for each metal layer for a small grid size. The effective density is then calculated using a moving average window across the chip (see Fig. 4.5), which is defined as the area of metal divided by the area of the window within a given region. The size of the window is denoted as the interaction distance or planarization length. The step or grid size to be used is determined by the planarization length to some extent. If the interaction distance is relatively large (e.g. 3-4 mm), the variation in pattern density is more gradual and a grid size of several hundred microns may be appropriate. If the interaction distance is less than 1 mm, a grid size of around 100 μm or smaller should be used. Differences in the planarization length are a function of the pad (e.g. soft vs. hard) and process type (e.g. metal vs. oxide CMP), as well as pad pressure, slurry type, and other process variables.

In this work, we use a square, equally weighted window to calculate effective density. Ouma has extended the CMP model to use an "elliptically" weighted circular symmetry planarization response function or filter [59, 60]. However, the errors from using a square window are relatively small and thus a square window is used in this thesis for simplicity.

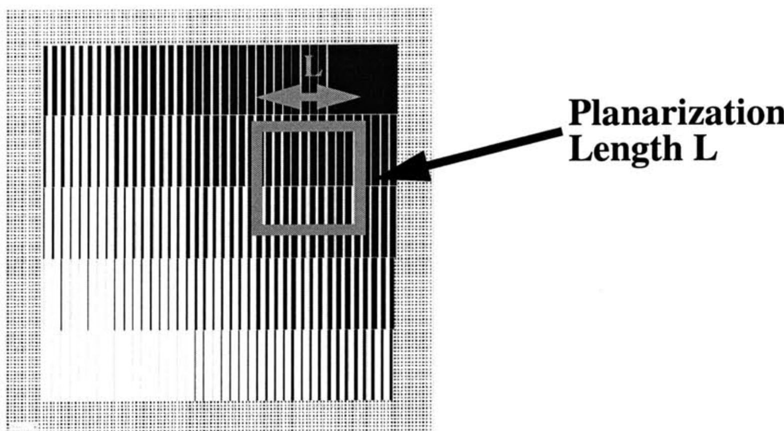
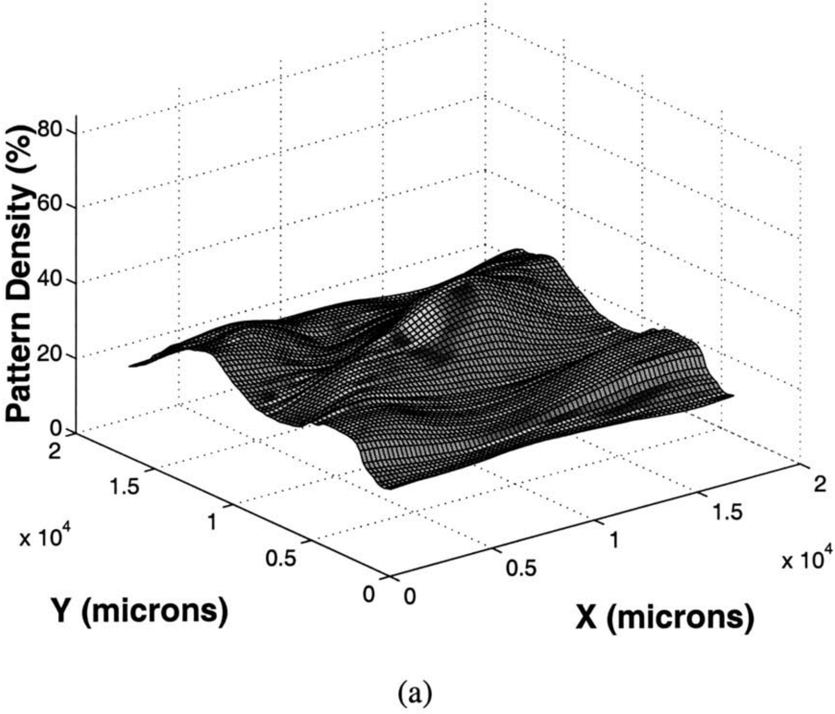


Figure 4.5: A moving average square window with planarization length L is used to compute the effective pattern density at different points on the die.

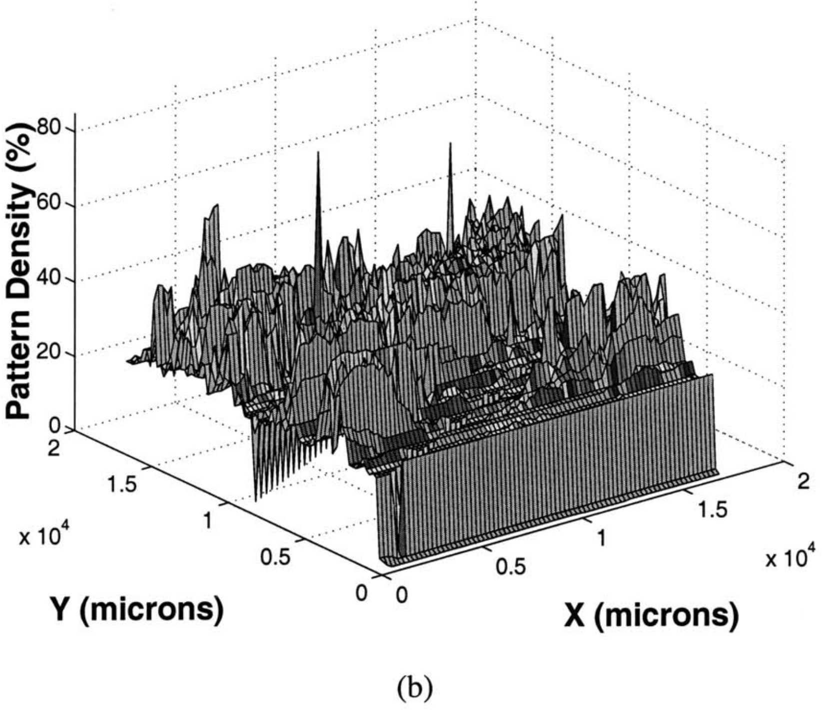
Fig. 4.6 shows the effective pattern density calculated for the M5 (Metal 5) layer in a 1 GHz microprocessor [61] with interaction distances of 3.5 mm and 100 μm. The large interaction distance results in a smoothly varying effective pattern density compared with the sharp spikes with a small interaction distance. The effective pattern density information is used with CMP models, with Fig. 4.6 (a) representative of the density used for oxide CMP and that in Fig. 4.6 (b) typically used with a metal CMP model.

The effective pattern density given in Fig. 4.6 (a) can be used to compute the variation of the ILD thickness between M5 and M6. Using values of 1 μm for the metal thickness and nominal ILD thickness for a target of 50% pattern density, the computed ILD thickness is shown in Fig. 4.7 (a). This thickness map allows us to model the variation as a function of spatial location within the chip. We obtain a minimum ILD thickness variation of -5.9% and a maximum ILD thickness variation of -28.2%. However, since this chip has lower than 50% average density, all points on the die are thinner than the nominal target ILD, and M5 to M6 capacitances will be substantially larger than designed. We can use the methodology described in the previous section to *predict* the ILD thickness variation before actually performing CMP. This allows us to reduce the maximum variation from the target ILD thickness by centering the process (see Fig. 4.7 (b)). Since the chip has relatively low pattern density throughout, the polish time can be adjusted (reduced) so that the dielectric is not unnecessarily overpolished. This results in a variation of -11.5% to +11.5% while the total variation range remains the same (22.3%).

An alternative is to increase the pattern density by adding metal (known as metal or dummy fill) to empty areas of the block [62, 63]. While very useful for chips with large blocks of varying density, a disadvantage of this technique is that some nets may suffer from added capacitance. This can cause a larger signal delay as well as increase crosstalk, and careful placement of the metal is necessary for effective use of fill.

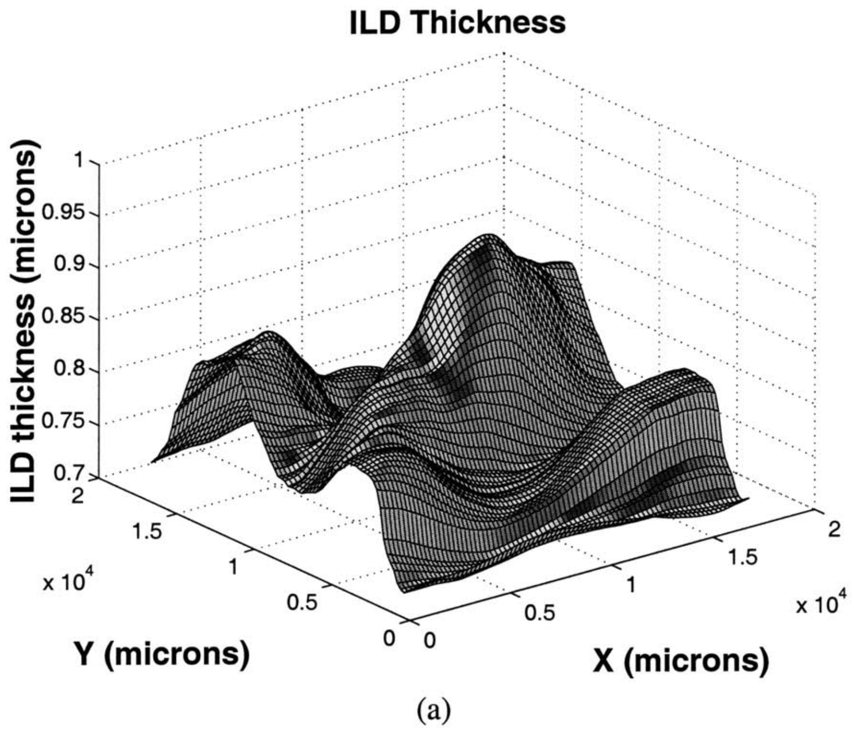


(a)

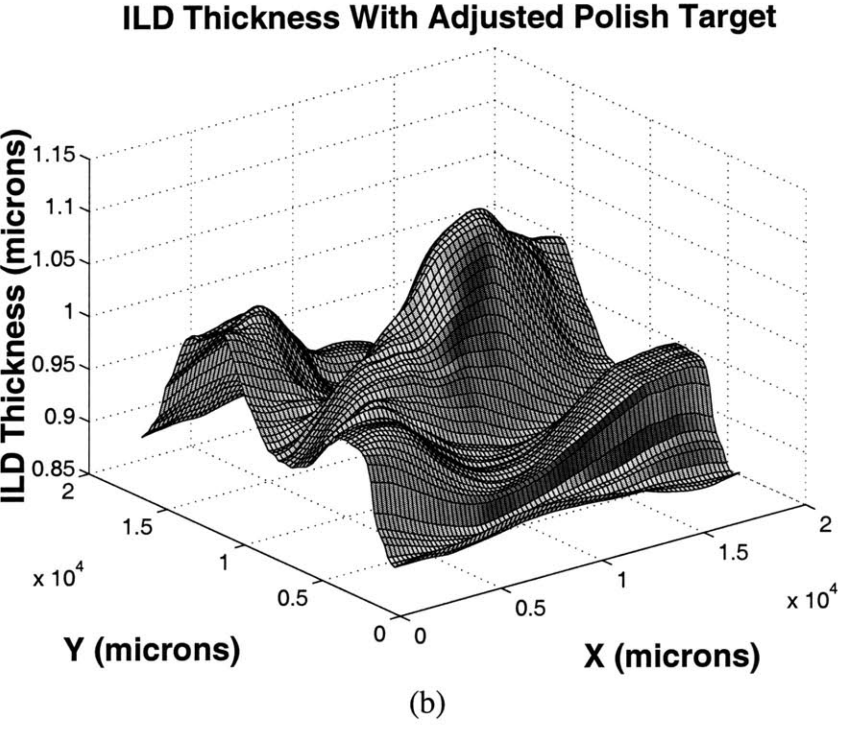


(b)

Figure 4.6: The effective pattern density for the Metal 5 layer of a microprocessor using an interaction distance of 3.5 mm (a) and 100 μm (b).



(a)



(b)

Figure 4.7: Simulated ILD thickness obtained by setting the polish time for a nominal target ILD thickness of 1 μm at 50% pattern density (a). The ILD thickness obtained by recentering the polish time based on average effective pattern density (b). The range of variation remains the same, while the maximum deviation from the 1 μm target is reduced.

Chapter 5

Systematic Variation Impact Assessment

This chapter studies the impact of variation on circuit performance. Specifically, we simulate the effects of variation on different test circuits to determine the impact on signal delay and clock skew. We look at three case studies, of which the first two are taken from high performance microprocessors. For the microprocessor case studies, we apply the systematic variation analysis methodology presented in the previous chapter to study the impact on signal delay and clock skew. The first case study analyzes the impact of ILD thickness variation on interconnect delay in a 1 GHz microprocessor fabricated in a 0.25 μm technology with aluminum interconnect. The second case study considers a different 1 GHz microprocessor that it is fabricated in a 0.18 μm technology and uses copper interconnect. In this case study, we consider the effects of metal thickness variation on clock skew resulting from CMP of the interconnect in a damascene process. The clock skew is also simulated assuming that aluminum interconnect is used instead. Additionally, we compare the effects of polysilicon critical dimension (CD) variation with that due to CMP variation. Finally, in the third case study we compare the effect of different processes (oxide vs. copper CMP) on interconnect delay in bus lines.

5.1 ILD Thickness Variation Impact on 1 GHz Microprocessor

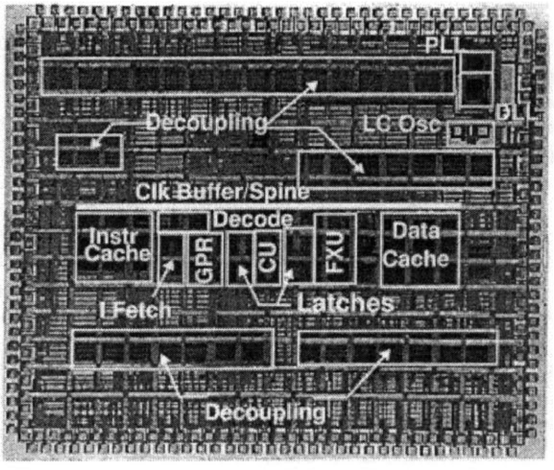
In our first case study, we analyze the impact of pattern dependent variation on global interconnect delay [48]. As discussed in Chapter 1, not having systematic parameter variation models, designers often use worst-case limits to bound the variation. This can lead to unnecessarily large design margins. Our first goal is to make use of the pattern dependent variation model for the ILD thickness and determine how much benefit one may gain from

such modeling. We determine the effects of ILD thickness variation on global path delays in a 1 GHz microprocessor [61], shown in Fig. 5.1. The percentage ILD thickness variation from nominal is computed based on the effective pattern density. We study the effect of variation on all the global paths in the chip for three different cases:

(i) Pattern dependent ILD thickness variation model,

(ii) ILD thickness variation assuming the addition of metal fill,

(iii) Worst-case ILD thickness variation across the chip.



Courtesy of IBM
Austin Research Lab

Figure 5.1: The 1 GHz microprocessor used in simulating the effects on global path delay due to ILD thickness variation.

The microprocessor is designed by IBM using six levels of metal with aluminum interconnect. The chip size is 7.4 mm x 8.1 mm, and it is designed to operate at a global clock frequency of 1 GHz. We simulate the impact of ILD thickness variation on interconnect capacitance and delay for all of the approximately 6200 global nets and 2100 global paths.

5.1.1 Computing ILD Thickness Variation

The ILD thickness variation is calculated using the pattern density information for all six metal levels and the model described in Chapter 2. The pattern density is first extracted

using an IBM density extractor for a grid size of 100 μm with a square window. A CMP planarization length of 3.5 mm is used to compute the effective pattern density for all metal layers on the chip. The ILD thickness variation is computed according to Eq. 3.6. Table 5.1 lists the resulting range of ILD thickness variation between all metal levels. The ILD thickness variation numbers are for the dielectric below the metal level stated in the table. The worst-case tolerances are taken from the design manual guidelines for this process. The effective pattern density is relatively low for all metal layers on this chip (less than 50%) using a planarization length of 3.5 mm, resulting in a thinner than nominal ILD thickness. Also, the range of pattern density variation is small, resulting in a narrow range of ILD thickness variation (e.g. -20% to -30% for metal layer 3). However, the worst-case design guidelines assume an ILD thickness variation of up to ±40 %.

Table 5.1: ILD Thickness Variation (%)

Level	Minimum	Maximum	Worst-Case Tolerance
M2	-17	-25	±30
M3	-20	-30	±40
M4	-17	-24	±40
M5	-17	-24	±40
M6	-16	-23	±40

5.1.2 Effect of Metal Fill

Metal fill is commonly used to reduce the variation in effective pattern density across a given metal layer. In turn, this leads to a reduction in the intra-die pattern dependent ILD thickness variation. Fig. 5.2 shows how this reduction in ILD thickness variation can be achieved and algorithms have been proposed to obtain acceptable metal fill [62, 63]. There

are some drawbacks, however. First, there may not be enough large blocks of empty areas to add metal fill in all places. This is not likely to be a problem for our case study since the effective pattern densities are very low across all metal levels, leaving enough room to add fill. Second, adding fill increases coupling capacitance among nets. Stine et al. [63] consider the effects of added capacitance due to metal fill. If done properly, the added intra-layer capacitance among adjacent wires on the same metal layer can be minimized. However, the inter-layer capacitance may still increase, depending on the configuration of metal wires above and below. Therefore, the benefits of improving the ILD thickness uniformity may be offset to some extent by the increase in capacitance, especially inter-layer capacitance, depending on the aggressiveness of the fill strategy.

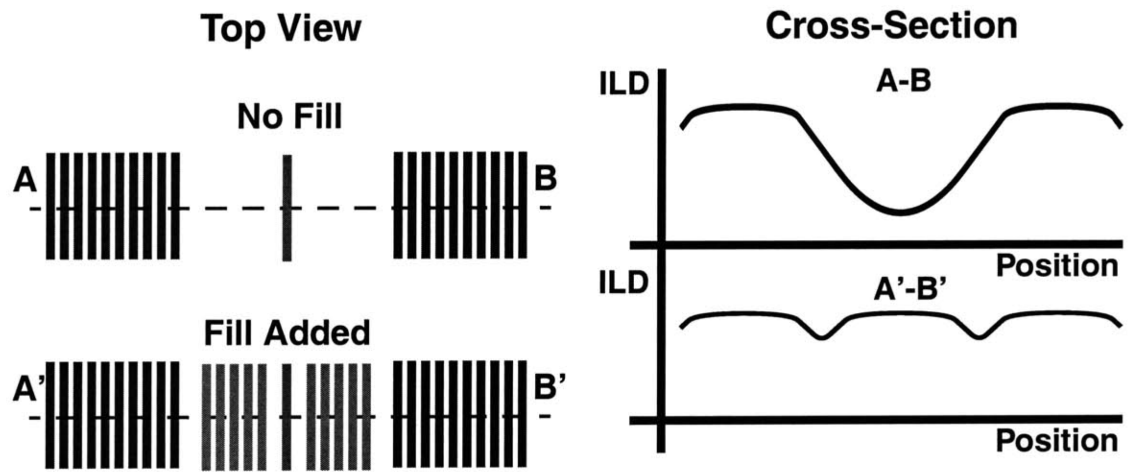


Figure 5.2: Adding metal fill reduces the ILD thickness non-uniformity by increasing the effective pattern density across the chip.

5.1.3 Interconnect Capacitance Distribution

The interconnect capacitance is modified for all global paths based on the ILD thickness as a function of spatial location. The effect of ILD thickness variation for all 6200 global nets is shown in Fig. 5.3 (a). We see that the net capacitance can increase by as much as 28% from nominal. For this case study, we find that CMP variation results in a thinner than nominal ILD thickness all across the chip. This is due to the fact that the effective pattern density is relatively low at all points on the die. If we assume a target polish time based on 50% pattern density, all areas on the chip will continue to polish even when the features have been cleared. This will cause a thinner ILD and increase the inter-layer capacitance. As discussed in the previous chapter, the process can be centered to reduce the overall variation.

For the case of metal fill, we assume that there is no increase in interconnect capacitance by adding fill. Given a minimum metal pattern density requirement of 30%, we look at the maximum potential benefit that can be obtained for the 6200 global nets by adding metal fill. Fig. 5.3 shows that the maximum increase in capacitance (from nominal) due to ILD thickness variation can be reduced from 28% to 18%.

5.1.4 Interconnect Delay Distribution

We next consider the simulated delay variation of all 2100 global paths in the microprocessor. Fig. 5.4 compares the delay distributions when the ILD thickness variation is modeled for the various cases (pattern dependent model, metal fill, and worst-case limits). Using a pattern dependent model for the ILD thickness variation results in a much tighter distribution than that predicted by the worst-case ILD thickness variation. Adding metal fill tightens the distribution even further. Since all metal layers on the chip have a thinner than nominal ILD thickness, there is an increase in the inter-layer capacitance, resulting in

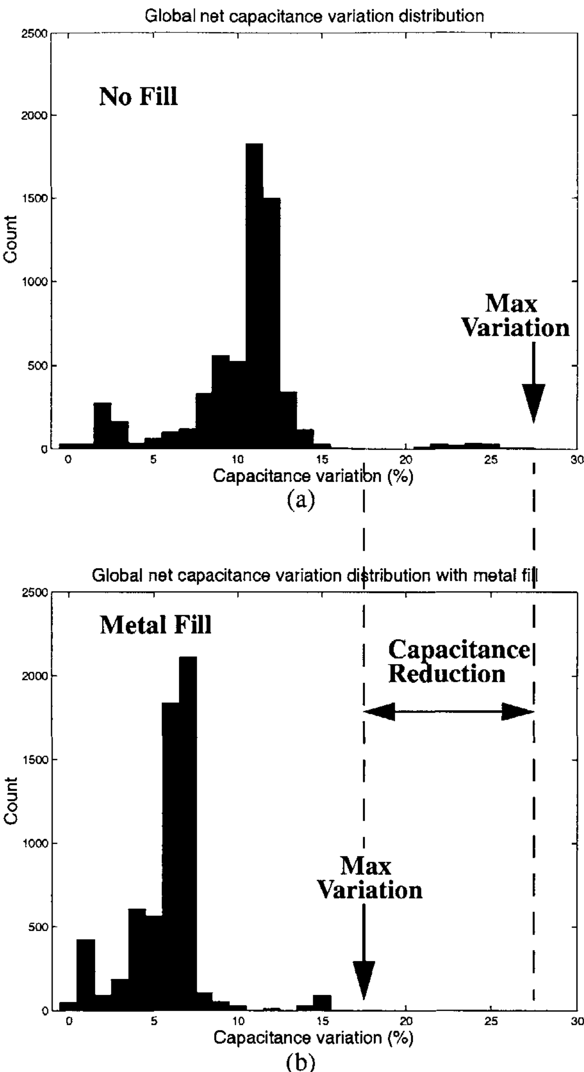


Figure 5.3: The global net capacitance distribution without (a) and with (b) metal fill. The addition of metal fill reduces the maximum capacitance variation from 28% to 18% (compared with nominal values).

an increase in the signal delays compared to the delays computed using the nominal ILD thicknesses.

Fig. 5.5 summarizes the results of this case study. We see that using the worst-case ILD thickness variation across the chip, the delay variation (increase) is as much as 88 psec compared with the nominal ILD thickness. Using the pattern dependent variation model, however, the maximum delay variation is only as much as 40 psec. This 48 psec

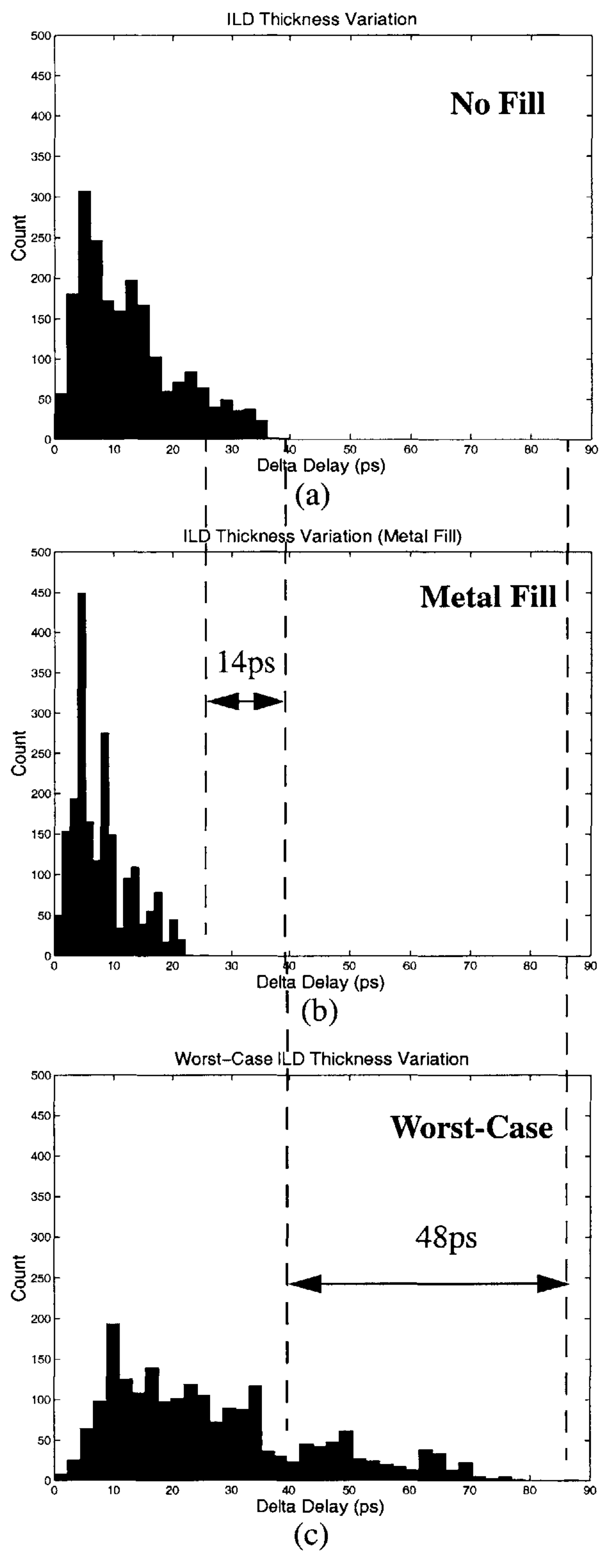


Figure 5.4: The global path delay variation for the microprocessor shown in Fig. 5.1. The delay variation is simulated with the pattern dependent ILD thickness variation model (a), assuming enough metal fill is added to limit the minimum pattern density at 30% (b), and the worst-case limits on ILD thickness (c).

reduction with the pattern dependent model corresponds to an extra design margin of about 5% of the clock frequency, and allows for approximately one extra stage of logic that may be introduced in the most sensitive paths. With technology scaling, this benefit may be an even larger percentage of the clock period. We also simulate the effect on delay assuming that metal fill is used to reduce the variation. In this case, the ILD thickness variation is calculated based on a minimum pattern density limit of 30%. As seen in Fig. 5.5, the effect of fill further reduces the delay variation compared with the worst-case ILD variation. While a large portion of the performance gain can be realized through a reduction in the delay uncertainty by simply *modeling* the across chip variation, this case illustrates that actual variation reduction efforts (in this case by metal fill) can be an effective approach when feasible.

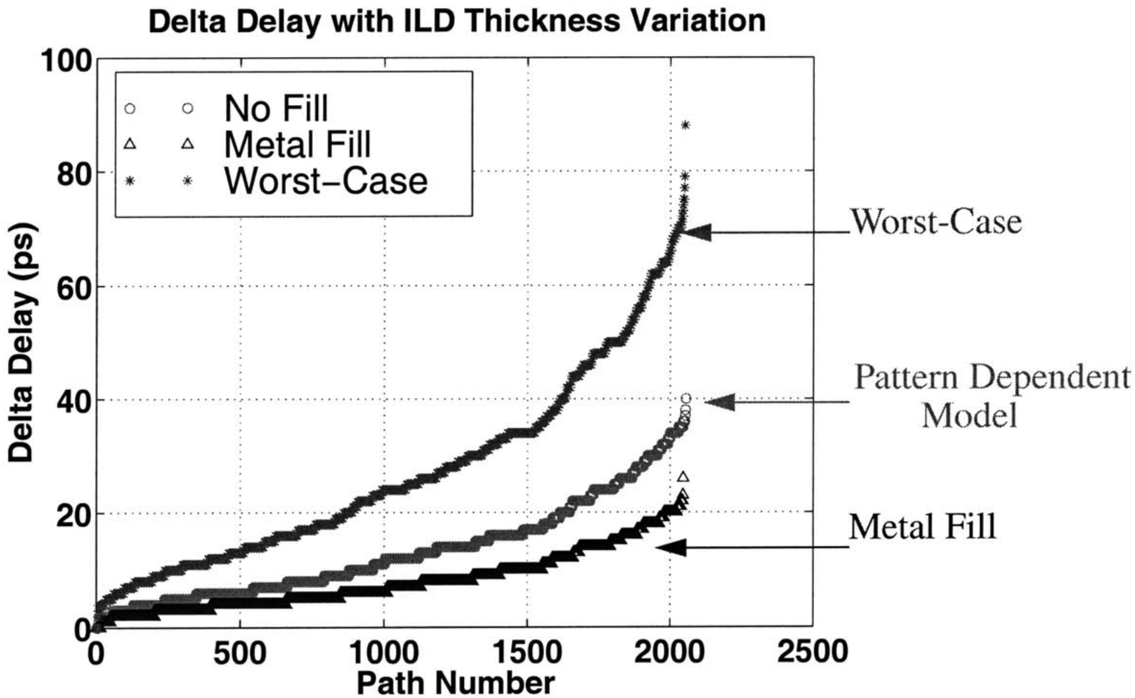


Figure 5.5: Summary of the interconnect delay variation for the various cases in Fig. 5.4.

5.2 Copper CMP and Poly CD Variation Impact on 1 GHz Clock Tree

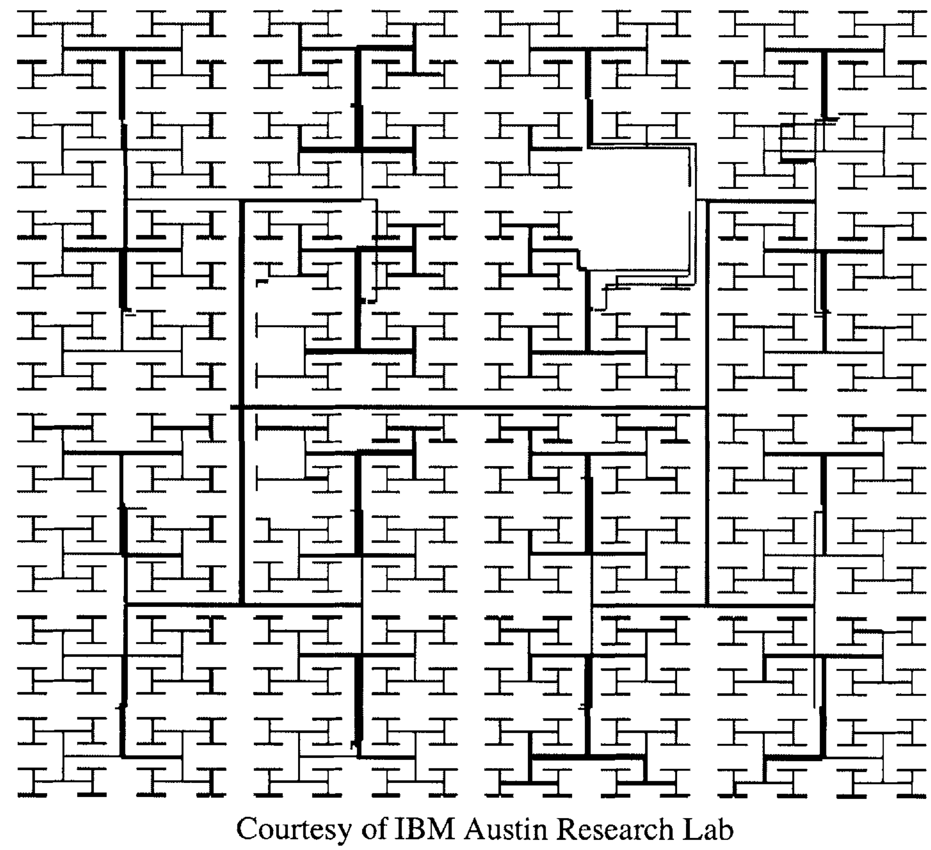
In the second case study, we study the effects of both interconnect and device variation on clock skew [49]. A nearly symmetric H-tree taken from a 1 GHz microprocessor [50] designed with *copper* interconnect is used as the test case (see Fig. 5.6). Unlike an oxide CMP process where the variation is in the ILD thickness, a metal CMP process suffers from metal thickness variation. This is due to two effects known as dishing and erosion [33], as shown in Fig. 3.4. Although the original design uses copper interconnect, for a comparison between technologies we also simulate the impact of variation assuming aluminum interconnect is used. In addition, the effect of poly CD variation is included to compare the impact among interconnect and device variation sources.

5.2.1 Interconnect Variation Modeling

Pattern dependent interconnect variation is modeled based on process data. Both dishing and erosion effects are considered for copper CMP. The ILD thickness variation is modeled using the model described in Chapter 3 for the aluminum interconnect case.

5.2.2 Poly CD Variation Modeling

The poly CD variation is based on the results from [37], where pattern dependent effects are measured for poly lines in the 0.35 μm generation. The data in [37] shows that differences in pattern density and pitch can affect the linewidth variation by as much as 15% or more. We simulate the device poly CD variation based on the configuration given in Fig. 5.7 to account for within-die gradient effects. The chip is divided into four quadrants, with a limit on the maximum poly linewidth variation of 5%. The assumption is that some initial compensation is performed using optical proximity correction (OPC) techniques. The clock skew is simulated using cascaded drivers at the input and loaded with latches at the output. All devices in the latches experience the same amount of poly CD



Courtesy of IBM Austin Research Lab

Figure 5.6: The clock tree used to simulate the effects of copper CMP variation on skew. The tree is driven by a series of cascaded drivers at the input and loaded with latches at the outputs.

variation based on the quadrant. Although the device variation is simulated in this way, the CAD tool implementation is general enough to handle spatial pattern dependence on every device, provided a poly CD variation model. We simulate the effects of variation sources on clock skew for five different cases:

(i) Cu interconnect -- no variation,

(ii) Al interconnect -- no variation,

(iii) Cu interconnect variation only,

(iv) Al interconnect variation only,

(v) Cu interconnect with poly CD variation.

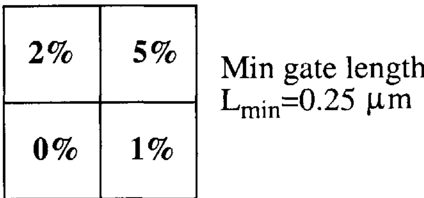


Figure 5.7: The configuration used to emulate the effects poly CD variation. The chip is divided into four quadrants, representing a fixed percentage of poly CD variation for all devices in that quadrant.

5.2.3 Clock Driver and Latches

The clock driver consists of a series of cascaded inverters (see Fig. 5.8), with increasing buffer sizes. The buffers are sized according to [3] with a total of *N* buffers so that the last buffer can drive the total interconnect load.

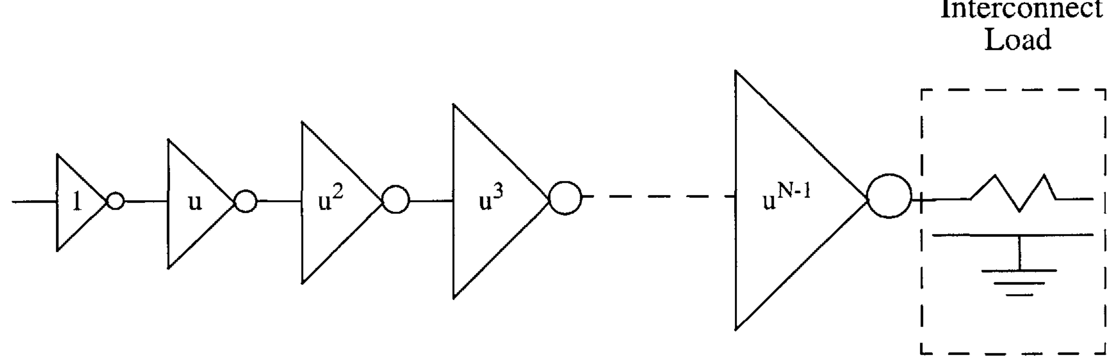


Figure 5.8: The clock driver consists of a series of inverters. The inverters are sized such that each successive buffer increases in size by a factor of u. The optimal value of u is equal to e=2.7182, and u=3 is taken as a good approximation.

We use the D flip-flop shown in Fig. 5.9, which consists of two cascaded T-latches. The flip-flop is positive edge triggered, so that the data latches at the output on the rising edge of the clock. To measure the skew, we set the data input to high and determine the difference in clock arrival times at the output of the latches. The functionality of the flip-flop (simply referred to as a latch or register from here on) is shown in Fig. 5.10.

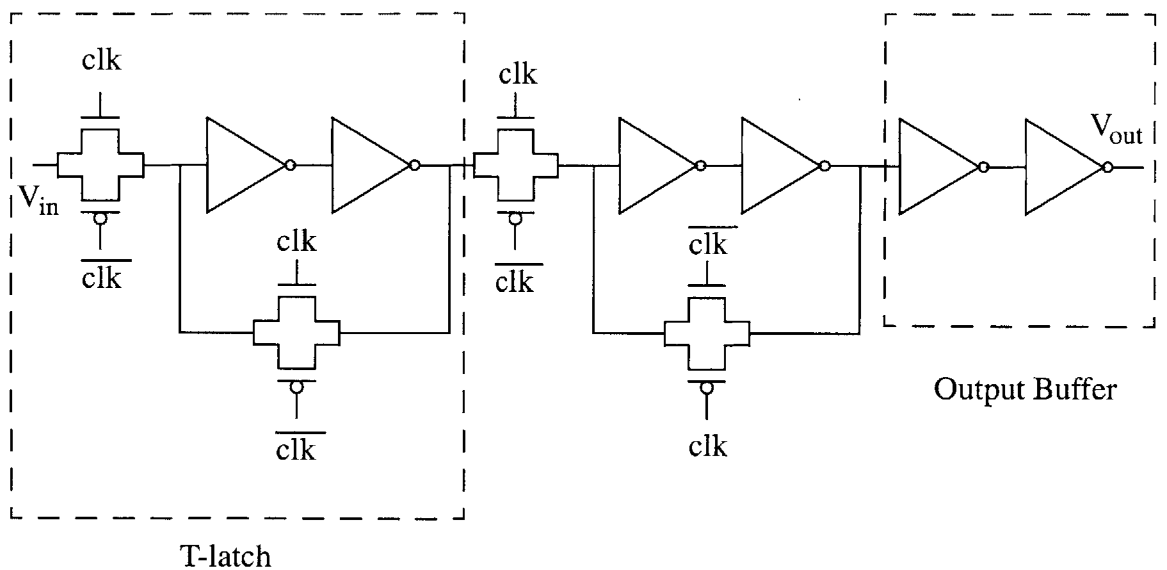


Figure 5.9: The D-flip flop (register) used at the outputs of the H-tree. The register is positive edge triggered, so the output latches at the rising edge of the clock. The input (V_{in}) is set to V_{DD} and the clock skew at V_{out} is due to variation in the interconnect (coming from clk) and the device variation in the register.

5.2.4 Clock Skew

We compare the maximum skew among all 996 paths in the H-tree to determine which source contributes the most skew. The simulation results are summarized in Table 5.2 and Fig. 5.11. This skew is due to both the asymmetry of the clock tree as well as the interconnect and device variation. First, note that even without any variation, a maximum clock skew of 34 psec exists (with copper interconnect) due to the asymmetry. When aluminum interconnect is used instead, the clock skew is 59 psec. Comparing the copper and aluminum cases with no variation, the copper interconnect results in less skew. This is expected since the total delay from the clock driver to the tip is smaller with copper because of the lower resistivity. Next, when CMP variation is added, both the copper and aluminum interconnect exhibit similar effects on skew. With copper CMP variation the skew increases to

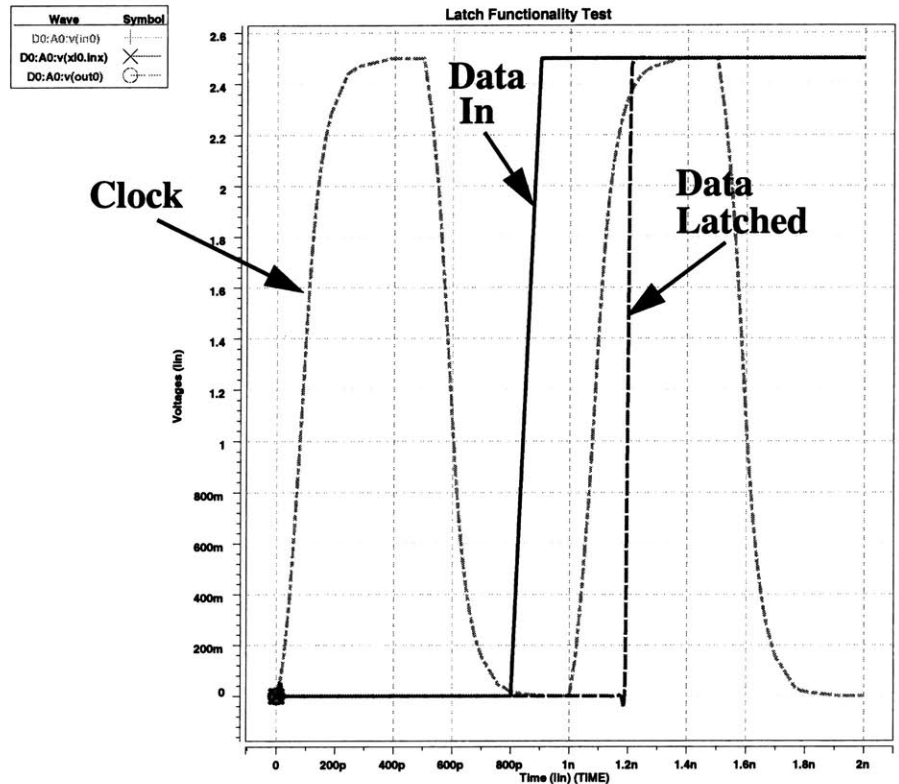


Figure 5.10: Functionality test of the latch. Data is set to high and arrives at the latch input before the clock goes high. The data is latched on the rising edge of the clock.

40 psec, and with oxide CMP variation (aluminum interconnect) the skew is 62 psec. There is only a small increase in skew with CMP variation, compared with the skew due to the asymmetry of the circuit. When poly CD variation is included, the maximum skew increases tremendously, even with a relatively small or conservative 5% variation. With copper interconnect, the maximum skew due to poly CD variation is 83 psec.

There is an important reason for the minimal impact on skew due to CMP variation. Although CMP variation changes the interconnect delay throughout the clock circuit, the skew or *difference* in delay does not change very much. The path delay in a tree is different from a single interconnect wire, since the common resistance of a given path must also charge other capacitors not in that path [2, 3]. Therefore, for a tree with around 1000 paths

as in the clock distribution circuit, a large difference in the interconnect resistance or capacitance is required at the tips of the clock tree to have an appreciable effect on the skew.

Table 5.2: Maximum Clock Skew

Interconnect	Spatial Variation Source	Max Skew (ps)
Cu	None	34
Cu	Metal Thickness	40
Al	None	59
Al	ILD Thickness	62
Cu	Device Poly CD	83

5.3 Interconnect Delay Variation: Metal vs. Oxide CMP

As we have seen, understanding the fundamental causes of variation are very important in effectively modeling the variation and determining its impact on performance. In this section, we compare the effect of variation in an oxide CMP process with that of a metal CMP process. The differences are due to both the process itself and the type of interconnection metal used for wiring. The process effect makes a difference because a damascene CMP process results in pattern (density, linewidth, linespace) dependent metal thickness variation compared with pattern (density) dependent ILD thickness variation in an oxide CMP case. Another process effect is the difference in the interaction distance used to compute the effective pattern density required by a CMP model. There can be an order of magnitude difference in the planarization lengths, depending on the type of the process, process conditions (e.g. slurry, pad), or the machine (e.g. details of CMP carrier and table design). Typically, long range interactions on the mm scale are required to model

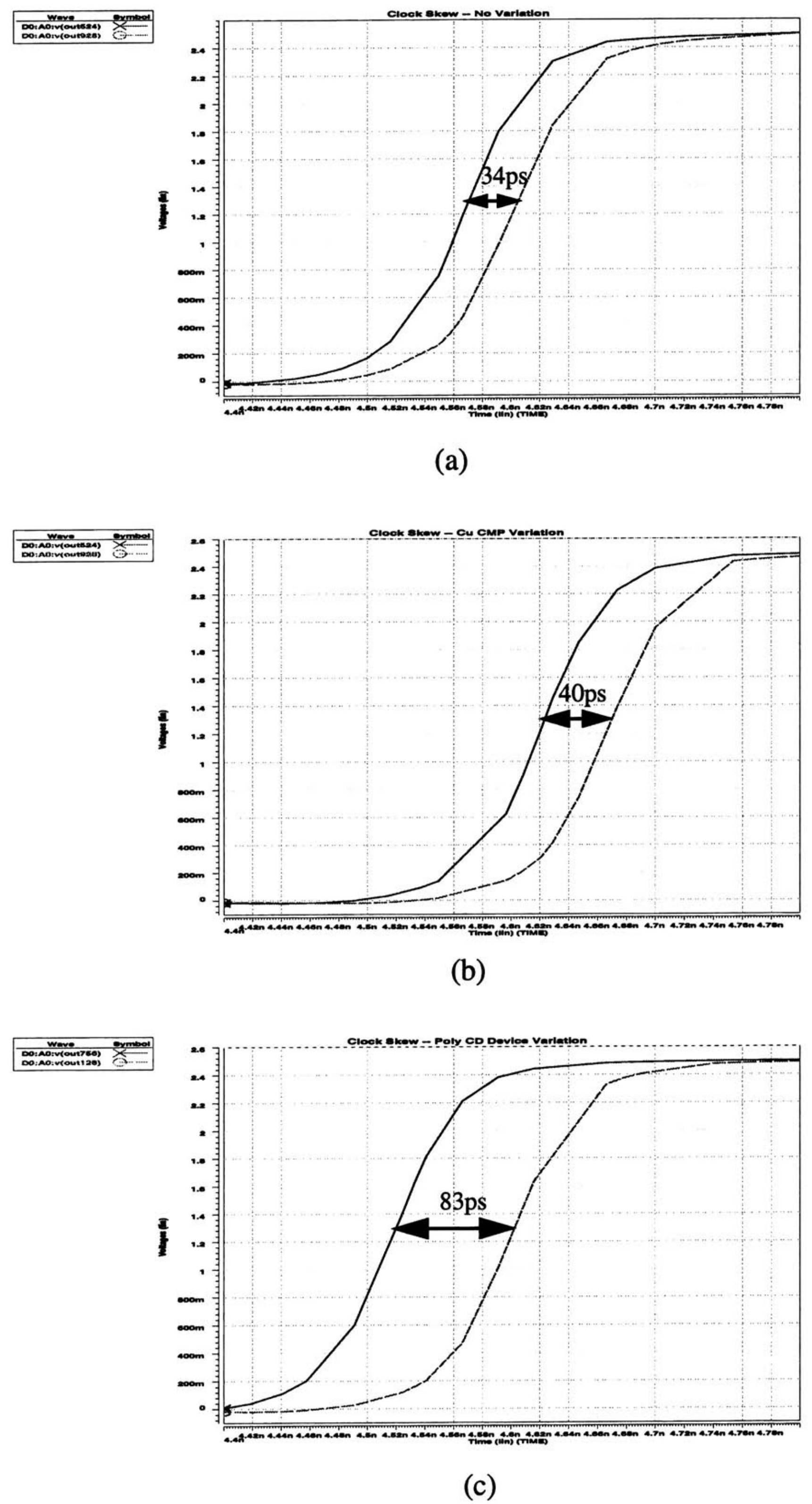


Figure 5.11: The delays of the fastest and slowest paths in the clock tree with no variation (a), copper CMP variation (b), and poly CD device variation (c).

the variation in an oxide CMP process compared with shorter range interactions of several hundred microns required for metal CMP. The type of interconnect used also makes a difference in the resistivity of the interconnect. Copper has a lower resistivity than aluminum and results in a lower interconnect delay. Therefore, the effect of variation on absolute delay is not as large in the copper case as it is with aluminum interconnect.

5.3.1 Copper CMP Interconnect Variation

The effect of CMP on interconnect delay is modeled on an array of lines such as those found in a data bus. The array is of relatively fine pitch, and erosion dominates. The metal thickness variation is a function of bit position within the array since short range interactions are important. Fig. 5.12 (a) shows that metal thickness varies at the edge of the array compared with bits toward the center, and a sample profilometry scan is given in Fig. 5.12 (b). Using an interaction distance of 100 μm (with blanket oxide surrounding the region immediately next to either edge of the array) and the erosion data for various pitch values, we simulate the effect of erosion on delay for a 128 bit array. For our case study, bits at the edge of the array result in *more* thickness loss than those near the center. Depending on the process, however, the erosion profile may be such that bits near the edge of the array experience *less* erosion than those at the edge.

5.3.2 Oxide CMP ILD Thickness Variation

Unlike metal damascene CMP, the effect of oxide CMP is due to long range interactions. The result is that the effect of ILD thickness variation is very gradual for an array of interconnect. Therefore, an array of lines in an oxide CMP process will not experience significant variation within the array, but arrays placed across different corners of the chip will be affected by the effective metal pattern density (see Fig. 5.13).

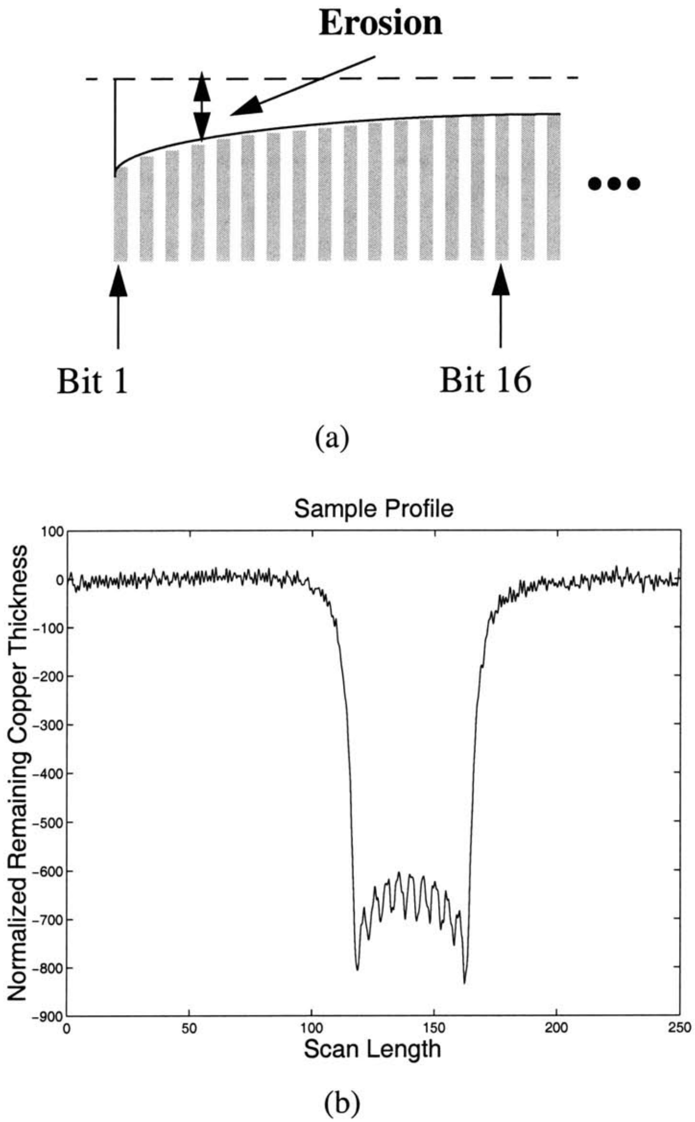


Figure 5.12: Amount of oxide erosion in an array of lines depends on the bit position within the array (a). A sample profilometry trace for an interconnect array (b) shows that for the CMP process conditions in this example, lines near the edge of the array experience more metal thickness loss than those near the center.

5.3.3 Interconnect Delay Variation

The interconnect delay variation for the 128 bit array is simulated for two different cases, with both copper and oxide CMP. The pitch is varied between minimum and 5X for interconnect in 0.25 μm technology. Fig. 5.14 shows the simulated effect on delay variation for the copper CMP case using data from the MIT/SEMATECH Cu Mask [39]. We

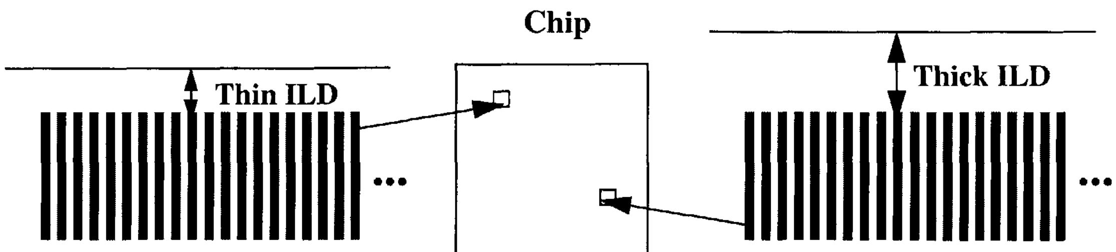


Figure 5.13: In an oxide CMP process, bit position within the array does not impact the ILD thickness variation. Long range interactions result in a variable ILD thickness for the entire array depending on the array location with the chip.

see that the minimum pitch lines experience the greatest amount of sensitivity to variation. With the given process conditions for this study, the lines near the edge of the array have a larger thickness loss than those near the center. At minimum pitch, bit 1 has more than 35% delay variation and bit 64 (center bit) has more than 15% delay variation compared to the nominal case.

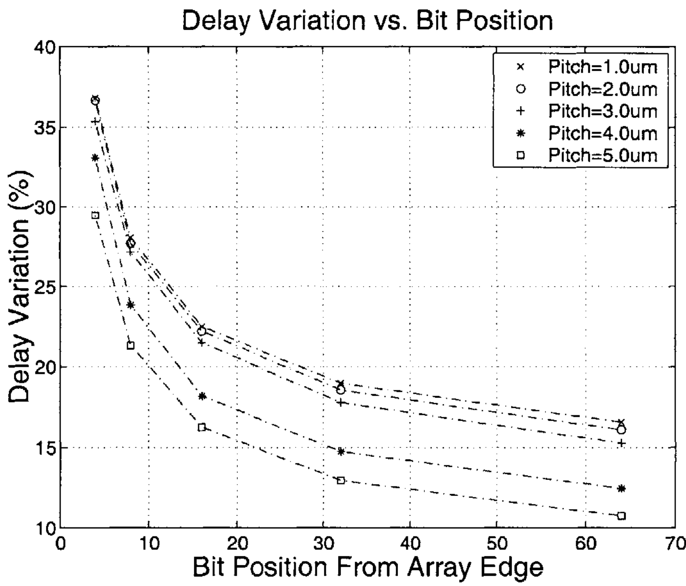


Figure 5.14: The simulated delay variation as a function of bit position for a 128 bit bus. A minimum size driver in 0.25 μm technology is used to drive each line, and the pitch is varied from 1 μm to 5 μm. Minimum pitch lines near the array edge are impacted the most after copper CMP.

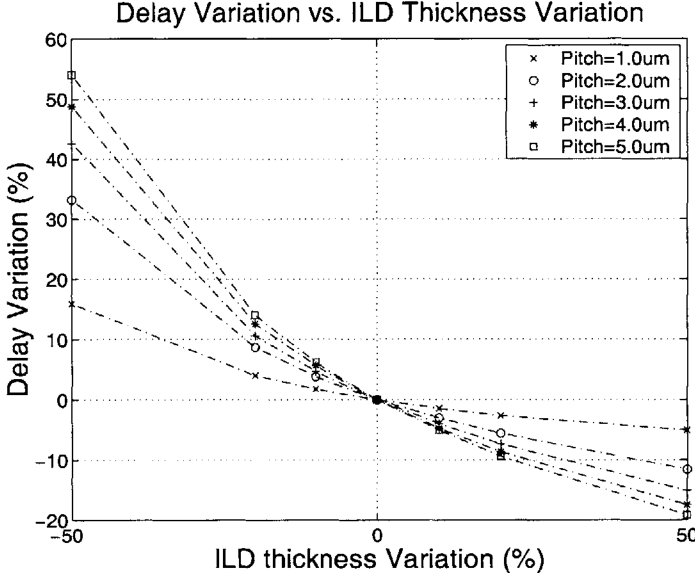


Figure 5.15: The delay variation for the entire bus as a function of ILD thickness variation for underlying metal pattern densities between 0-100%. The nominal ILD thickness is targeted at 50% density.

When considering ILD thickness variation, across chip pattern density should be taken into account. Although there is very little difference in delay as a function of bit position for an array of fine pitch lines, different arrays experience different amounts of variation. Fig. 5.15 shows the sensitivity to ILD thickness variation based on a pattern density variation between -50% and 50%. This represents the maximum worst-case variation one can expect in delay, even if there are areas of very high and very low density on the chip. We see that the effect of ILD thickness variation can either increase or decrease the delay, with a range of variation around -20% to over 50% at minimum pitch.

5.3.4 Discussion of Interconnect Variation Impact on Delay

Comparing the two processes (metal and oxide CMP), we find some important differences in the impact on circuit delay and the implications for variation modeling. First, in a metal (copper) CMP process, it is much more important to model the effects of CMP. With copper CMP, the variation is localized and differences in patterns across the chip must be

taken into account. Comparing Figs. 5.14 and 5.15 we see that the delay increase may be even greater with oxide CMP compared with copper CMP. However, this is not very likely since a microprocessor (see Table 5.1) will have a more limited range in effective pattern density (e.g. perhaps only ±20 %). This means that based on the range of pattern density, the process polish time can be centered to reduce the variation in ILD thickness. In a metal CMP process, that is not possible because there are always areas of (close to) 0% and 100% pattern density (see Fig. 4.6). Also, additional effects such as linewidth and linespace affect metal thickness variation.

Another difference is that with oxide CMP very high density regions tend to be beneficial since there may be a reduction in the delay due to a greater than expected ILD thickness. With copper CMP, metal thickness loss leads to an increase in resistance even though the lateral coupling capacitance decreases. This generally results in a greater delay because resistance tends to dominate, making it more important to model CMP variation.

The above results and discussion imply that CMP modeling is very important for accurate delay simulation and prediction, especially with copper CMP. Even when a complete density map is not available in the early stages of design, approximate densities should be used for better prediction of potential delay variation problems.

5.4 Summary

In this chapter, we have seen that interconnect variation can have a significant impact on signal delay. We have also seen that device poly CD variation impacts clock skew, even though the clock tree is interconnect dominant. In the following chapter, the effects of other variation sources will also be included. We will compare the effects of systematic vs. random variation sources, including device parameter variations such as threshold voltage and gate oxide thickness. Additionally, we also consider power supply variation.

Chapter 6

Systematic and Random Variation Effects

In addition to modeling the systematic variation, it is important to look at random variation effects also. In other words, once the variation components have been separated and the systematic components are modeled, we must consider these in the context of the total variation impact on the circuit. Therefore, we include both systematic and random variation components to get a better perspective on the impact of variation as a whole. In particular, we look at systematic variation due to CMP in the interconnect, which affects either metal or ILD thickness depending on the type of process. We model the metal linewidth (CD) variation as having a systematic as well as a random component. The device variation (V_t, t_{ox}) is mostly random, except for L_{eff}, which has both systematic and random variation components similar to metal CD. Finally, we consider the effect of random power supply variation.

We take another look at variation impact on clock skew and signal delay as well as the impact on crosstalk noise. The H-tree from the 1 GHz microprocessor is considered again, this time with intermediate buffers inserted to study the impact of variation on both devices and interconnect. For variation impact on signal delay, the interconnect is first optimized for a specified clock frequency target based on the SIA Roadmap. The maximum interconnect length, buffer size, and buffer number are found for different interconnect pitch in the local, intermediate, and global tiers. Then, variation in the interconnect and devices is included to find new optimized values. Finally, we look at the impact of variation on crosstalk noise as a function of interconnect length for different driver sizes, rise times, and wiring pitch.

6.1 Clock Skew in H-Tree

The clock tree shown in Fig. 5.6 is modified to include two stages of intermediate buffers or repeaters as shown in Fig. 6.1. First, the H-tree is simulated without any interconnect or device variation. The device parameters are taken from the SIA Roadmap and the worst-case variation is based on [64, 65], as shown in Table 6.1. The 3σ variation in t_{ox}, V_t, and V_{DD} is assumed to be 10%, but for L_{eff} is it assumed to be 20%. We simulate the effects of both interconnect and device variation using Hspice to compute the clock skew as described below.

6.1.1 Metal Thickness Variation

The metal thickness variation in the interconnect is modeled based on [34, 39] and the methodology from Chapter 4 is used to model the systematic variation based on spatial effects in the layout and CMP process conditions. As before, we do not include the effects of metal CD variation since clock wires are much wider than minimum pitch.

6.1.2 Device and Power Supply Variation

Unlike interconnect variation in the clock tree, device variation includes large systematic and random components. The device variation is computed separately for the buffer stages and the latch by modeling the H-tree with the appropriate interconnect load as given by [2] and shown in Fig. 6.2. The effects of device and power supply variation are modeled using an equivalent circuit containing a series of distributed RC interconnect sections. When a fork is encountered in the tree, the equivalent resistance is reduced by a factor of two and the capacitance doubles. This procedure is repeated each time there is a fork in the tree until a buffer stage is encountered.

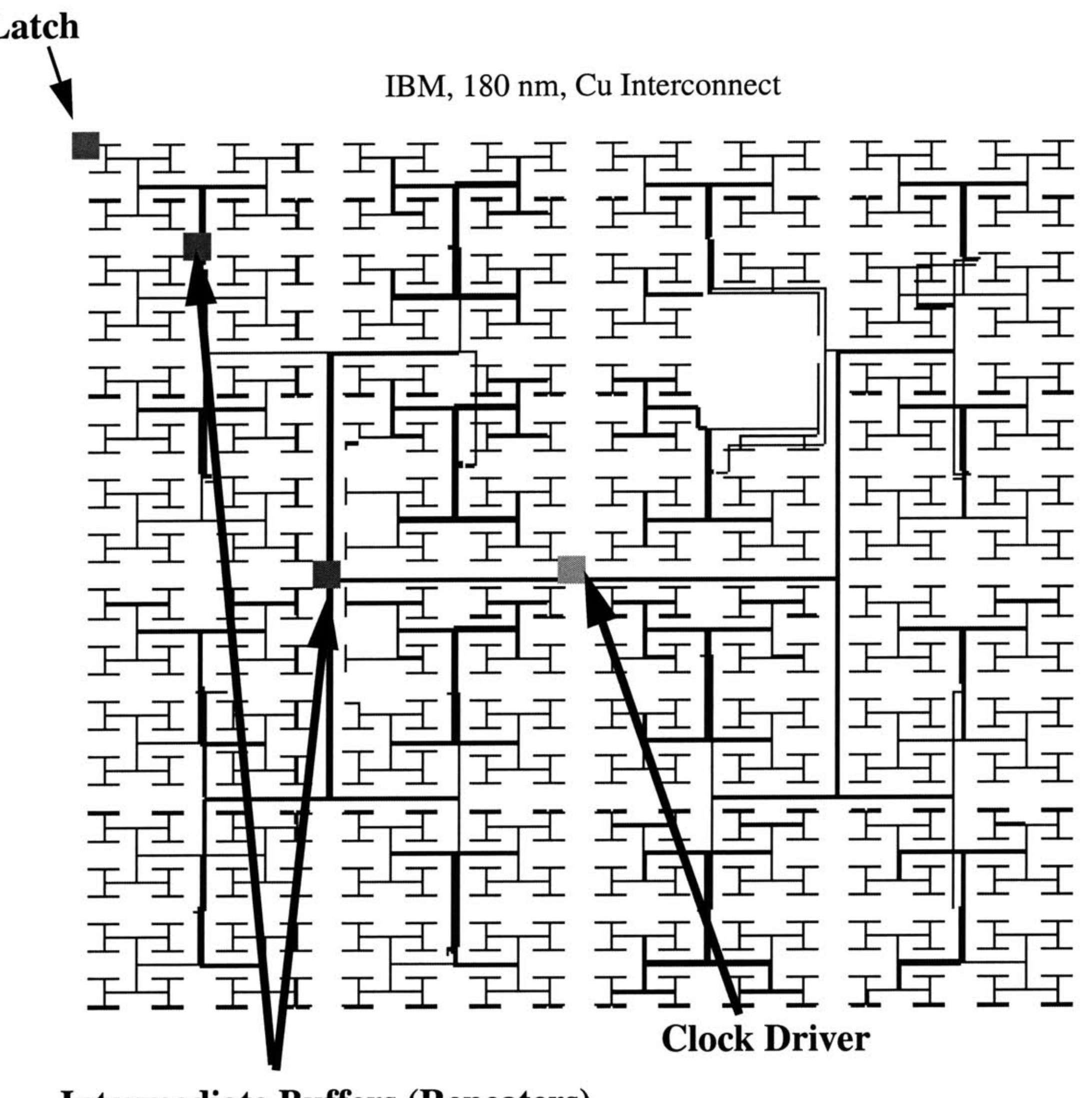


Figure 6.1: The 1 GHz H-tree is modified to include two stages of intermediate buffers. For clarity, the buffers are shown for only one path in the tree.

The power supply (V_{DD}) variation is considered random. Monte-Carlo simulations (1000 trials) are performed using the 3σ device and power supply variation tolerances given in Table 6.1 for two different cases:

Case 1: Use random values for L_{eff}, V_t, t_{ox}, and V_{DD} variation (within 3σ limits).

Case 2: Use random values for V_t, t_{ox}, and V_{DD} variation (within 3σ limits) and use 50% tighter tolerance for L_{eff} assuming fraction of variation is systematic.

Case 1 is referred to as the worst-case design approach, where the variation sources are assumed to be random. We refer to case 2 as the "tighter tolerance design" method.

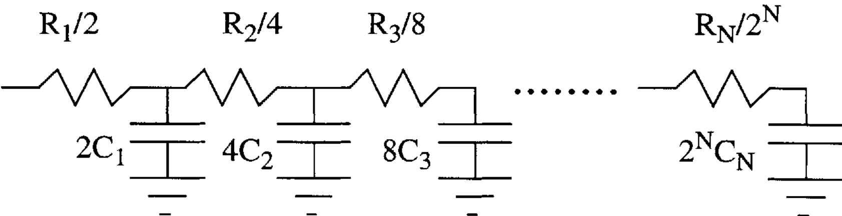


Figure 6.2: The equivalent RC network for a symmetric H-tree from the driver to the next buffer. When a fork is encountered, the resistance of the branch is reduced by a factor of two and the capacitance doubles.

Table 6.1: Device and Power Supply Parameters and 3σ Variation

Parameter	Nominal	Variation (%)
L_{eff} (nm)	180	20
t_{ox} (nm)	2.2	10
V_t (V)	0.40	10
V_{DD} (V)	1.8	10

6.1.3 Clock Skew

Table 6.2 lists the total clock skew for the various cases. There are several interesting points to note from these results. First, modeling the CMP variation in this circuit results in a skew reduction that is about 10 psec or 1% of the total clock cycle. A reduction in the L_{eff} tolerance, however, produces a skew reduction of twice this amount (about 20 psec). This shows that although the clock tree is an "interconnect dominant" circuit, it is still important to model the device variation. Second, compared to the systematic variation effects, random variations still account for a large fraction (almost 10%) of the clock

cycle. Finally, in this case study, the CMP variation results in lower skew than the case without any variation due to cancellation effects.

For the buffered H-tree, we find that systematic variation modeling predicts a lower skew than the case without any variation because we are able to take advantage of the asymmetry in the circuit. Here, variation impacts the delays of all the paths in the tree but some paths more than others depending on the neighbor surroundings. Note that if the tree were completely symmetric, the impact of variation would always result in a larger skew than the case without any variation (zero skew). However, a worst-case analysis would still result in a larger estimate of the clock skew than the tighter tolerance design method.

Table 6.2: Clock Skew in H-Tree With Interconnect, Device, and V_{DD} Variation

Variation Source	**Skew (psec)**	**Skew (% of clock period)**
None	45.53	4.55
Cu CMP (systematic model)	36.85	3.69
Devices & V_{DD}	116.58	11.66
Devices & V_{DD} (tighter L_{eff} tolerance)	96.96	9.70
Total Skew (worst-case device & V_{DD} tolerances)	162.11	16.21
Total Skew (systematic model for metal thickness & tighter L_{eff} tolerance)	133.81	13.38
Skew Reduction (systematic models)	**28.30**	**2.83**

6.2 Optimally Buffered Interconnect

When routing global wires over long distances, intermediate buffers are often inserted to minimize the path delay [2, 3, 66-68]. We describe here the methods from [2, 3] used to calculate the minimal interconnect delay and the optimal number of buffers that must be inserted in the circuit (see Fig. 6.3). We then consider the impact of variation on optimally buffered interconnect circuits.

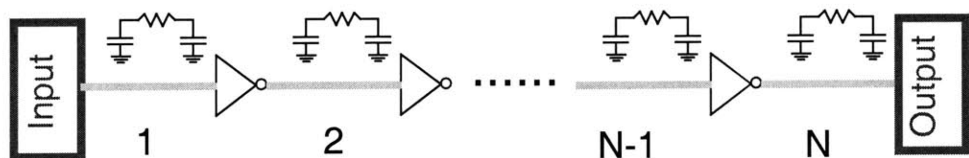


Figure 6.3: Intermediate buffers are inserted to minimize delay when routing a signal over long distances.

6.2.1 Buffer Insertion and Optimal Interconnect Delay

Buffer insertion can be used to reduce the delay of long wires. The insertion of buffers (or repeaters) reduces the dependence of interconnect length on delay from quadratic to linear, with the extra overhead of the intermediate buffer delays. The signal delay can be calculated to first order using the Elmore delay expression given by Bakoglu [2] as

$$Delay = \frac{0.4rcl^2}{N} + (N-1)t_{buf} + t_{latch} + t_{setup} \tag{6.1}$$

where

r = resistance per unit length,

c = capacitance per unit length,

l = interconnect length,

N = number of interconnect sections,

t_{buf} = buffer delay,

t_{latch} = latch delay,

t_{setup} = latch setup time.

The optimal number of buffers can be found by setting the derivative of the delay equation with respect to N to zero and solving for the optimal number of interconnect sections, N. The optimal number of buffers, *bufopt,* is equal to N-1.

Using Eq. 6.1, the optimal number of buffers can be found. However, we need to account for some additional effects. The buffer delay, t_{buf}, depends on the interconnect load and the driver size. In the above equation, this value is fixed. Also, in addition to the buffer number, the buffer size can also be optimized. To optimize both the number and size of the buffers, Eq. 6.1 can be modified as

$$Delay = \frac{0.4RC}{N} + 0.7(R_{tr}C + RC_L + R_{tr}C_LN) \qquad (6.2)$$

where

R = lumped interconnect resistance,

C = lumped interconnect capacitance,

R_{tr} = transistor on-resistance,

C_L = device load capacitance.

We represent the driver as just another buffer, so that there are a total of N equivalent sections containing a single input buffer with the total load equal to the interconnect plus the output capacitance of the driver and input capacitance of the next buffer. Increasing the buffer size reduces transistor resistance but increases the capacitance by the same factor. Setting h to be the optimal buffer size gives

$$R_{tr} = R_0/h, \qquad (6.3)$$

$$C_L = hC_0 \qquad (6.4)$$

with

R_0 = on-resistance of minimum size NMOS (or PMOS) transistor,

C_0 = capacitance of minimum size inverter (gate and drain capacitances).

Substituting into Eq. 6.2, we get

$$Delay = \frac{0.4RC}{N} + 0.7\left(\frac{R_0C}{h} + RC_0h + R_0C_0N\right). \qquad (6.5)$$

The optimal number of interconnect sections and buffers is then found by setting

$$N_{opt} = \sqrt{\frac{0.4RC}{0.7R_0C_0}} \qquad (6.6)$$

and the optimal buffer size is given as

$$h_{opt} = \sqrt{\frac{R_0C}{RC_0}}. \qquad (6.7)$$

6.2.2 Interconnect Length Variation

We now consider the effects of variation in optimally buffered interconnect. We choose a desired clock frequency constraint and compute the impact of variation on the maximum interconnect length, L_{max}. The metal linewidth variation behavior is similar to that of poly CD, where a fraction of the total variation is systematic. The metal thickness, device, and power supply variation are considered in the same way as for the clock tree. Interconnect parameters for a 250 nm process are used as defined in Table 6.3. The minimum wiring pitch for local, intermediate, and global wires is 0.5, 1.0, and 1.5 μm, with equivalent linewidth and linespace. The nominal metal and ILD thickness are also taken to

be equivalent. We assume that the 3σ variation tolerance for metal linewidth and thickness is 20%, 15%, and 10% for local, intermediate, and global interconnect, respectively.

Table 6.3: Interconnect Parameters for 250 nm Technology

Parameter	Value	Units
Global Clock Frequency	1.0	GHz
Dielectric Constant	3.5	--
Metal (Cu) Resistivity	2.2	$\mu\Omega$-cm
Local Wiring Pitch	0.50	μm
Intermediate Wiring Pitch	1.00	μm
Global Wiring Pitch	1.50	μm
Local Aspect Ratio	2.0	--
Intermediate Aspect Ratio	2.5	--
Global Aspect Ratio	2.67	--

Fig. 6.4 shows the degradation in L_{max} as a function of pitch, where the linewidth is held constant and linespace is varied from minimum (1X) to 3X for local, intermediate, and global interconnect tiers. The effect on L_{max} using both worst-case tolerances as well as a systematic model for metal thickness and a tighter tolerance on metal and poly CD variation are shown (assuming OPC is used). In addition, random variations in the buffers and the power supply are considered in the same way as for the clock tree, assuming that the variation sources considered are completely random for the first case and that a fraction of the poly CD variation is systematic for the second case. We define the degradation in L_{max} to be the necessary reduction in the maximum wire length to meet the clock frequency constraint. Using worst-case limits results in a degradation in L_{max} of 20% to more than 30%, depending on the wiring pitch and interconnect tier. As expected, the global lines are impacted the least. When copper CMP models and tighter design tolerances are

included, the result is a reduction in the L_{max} degradation, with a maximum degradation of less than 20%.

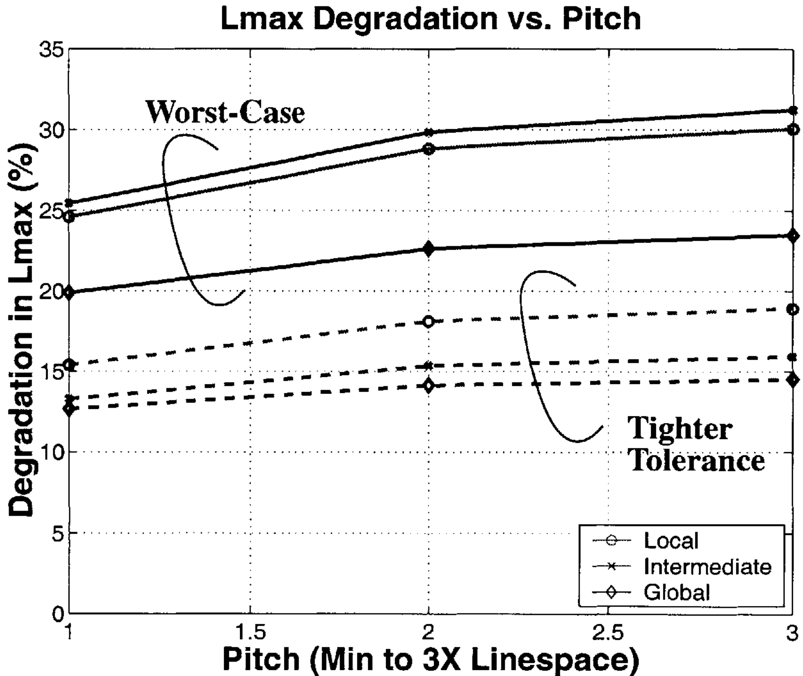


Figure 6.4: Degradation in the maximum interconnect length as a function of pitch for wires in the local, intermediate, and global tiers with worst-case variation tolerances as well as systematic variation models and tighter design tolerances.

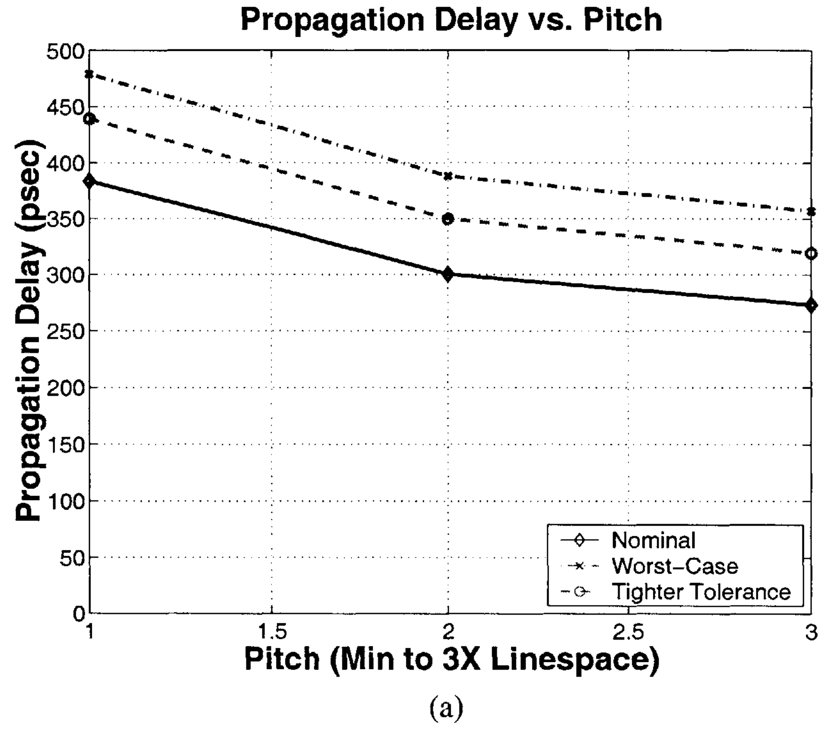
In this section, we have seen the impact of variation on interconnect length given a fixed clock frequency. Although variation can significantly impact the maximum wire length in optimized interconnect, we must compare this value with the chip side length. In Chapter 1 we showed that in the 250 nm generation, the maximum distance that can be routed using global and intermediate wires at minimum pitch is greater than two times the chip side length (see Fig. 1.16). In this case, optimal buffer insertion allows us to communicate between the furthest points on the chip in one clock cycle. Even if one clock cycle is enough to route between two points on a chip, it is still useful to optimize the interconnect. An optimized buffer insertion strategy ensures that there is more time available in the

clock cycle to perform other functions. If the maximum available wire length is greater than the maximum distance between the furthest points on the chip, the goal should be to minimize the delay for a fixed wire length. The next section uses the optimal buffer insertion technique to minimize the signal delay with and without the effects of variation.

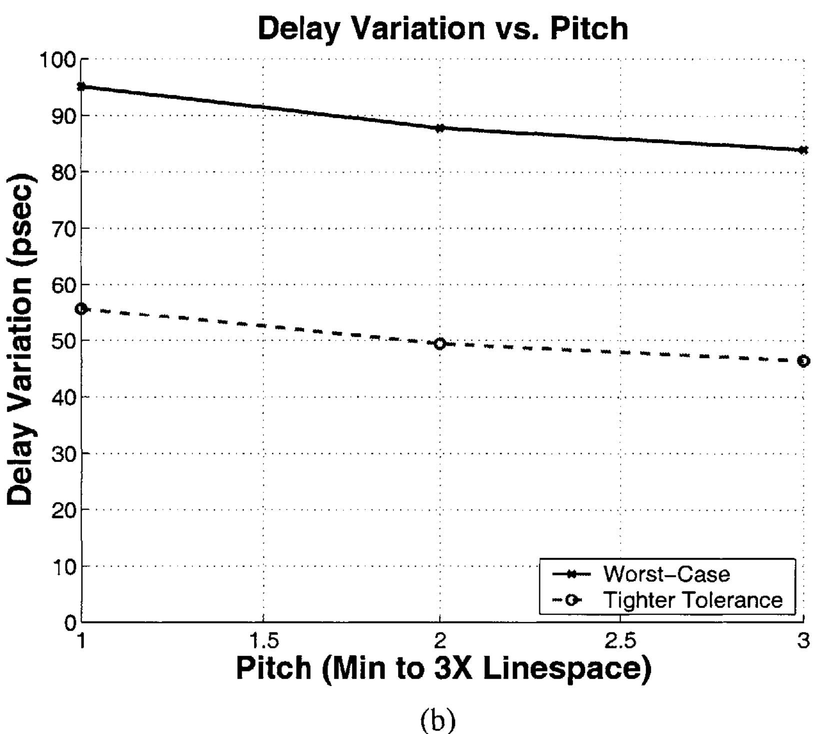
6.2.3 Delay Variation

To find the impact of variation on delay in optimally buffered interconnect, the maximum interconnect length is set to two times the chip side length. For the 250 nm generation, the chip side length is about 16 mm from the SIA Roadmap, so our study optimizes the delay based on an interconnect length of 32 mm. The nominal propagation delay and delay variation are shown in Fig. 6.5 for global interconnect. The signal delay at minimum pitch is 384 psec. Using worst-case limits for the variation sources considered here, a design margin of 95 psec must be added. This gives us a design guideline of 479 psec for the delay. With systematic modeling, this value is reduced to 440 psec. If this value is still not acceptable, the pitch may be increased. At 3X minimum linespace, the worst-case tolerance results in a design rule of 357 psec. However, using tighter design tolerances through systematic models, we obtain a delay specification of 350 psec at 2X minimum linespace. By reducing the linespace from 3X to 2X, the wiring pitch (*LW+LS*) is reduced from 4X to 3X, with linewidth equal to linespace. In this case, we are able to achieve a more aggressive design by reducing the global wiring pitch for this path by 25%.

The delay variation due to different sources independently is given in Fig. 6.6. We find that variation in the buffers accounts for most of the delay variation impact. Fig. 6.6 (a) shows that at 3X minimum linespace, the increase in delay using worst-case device and power supply variation tolerances is calculated to be 60 psec, compared to about 10 to 15 psec with metal thickness and linewidth variation. Note that as pitch decreases, the impact

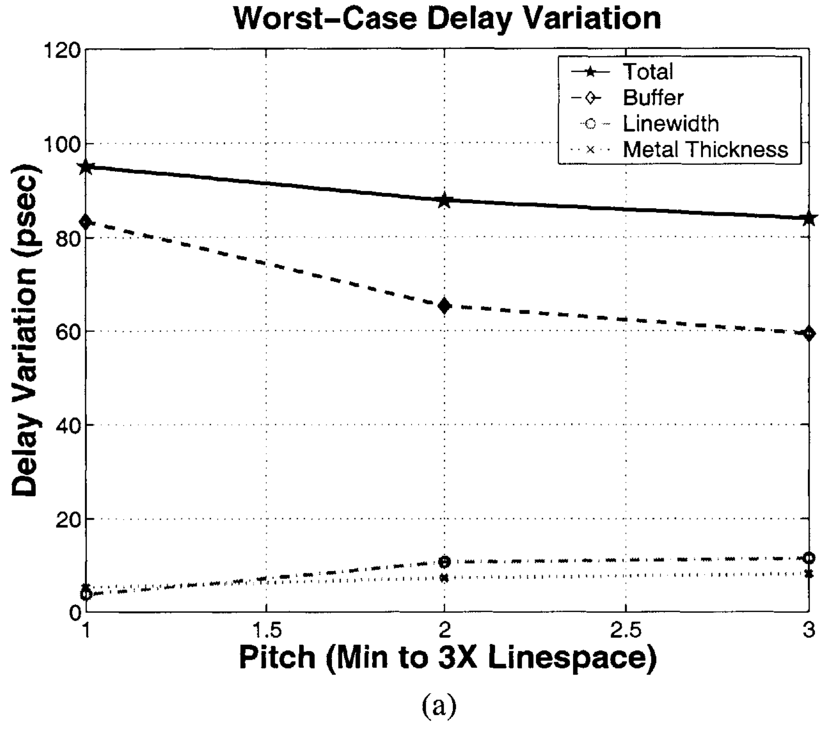


(a)

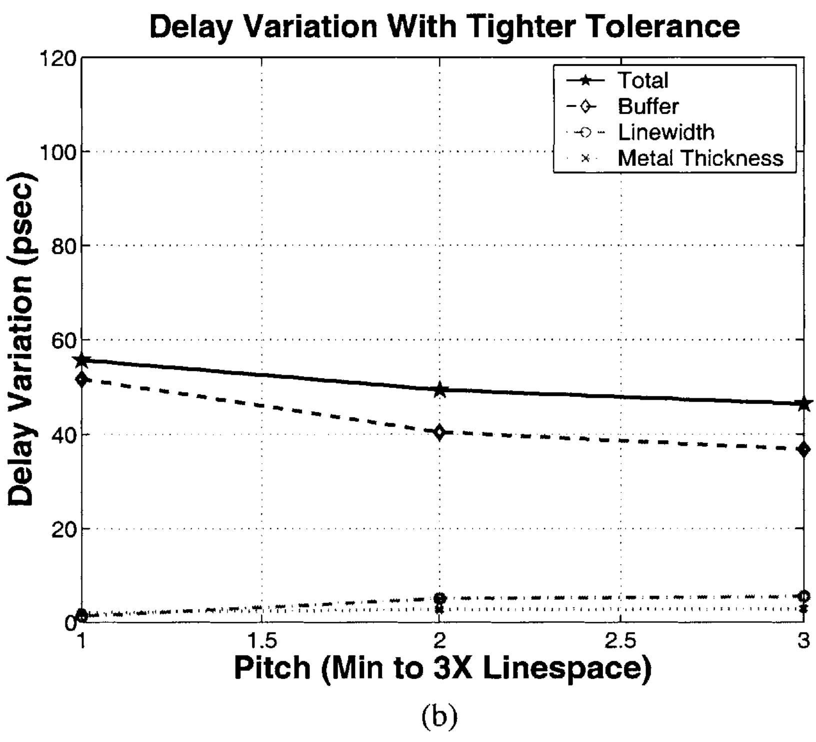


(b)

Figure 6.5: Propagation delay for optimally buffered interconnect (a). Delay variation with worst-case variation tolerances and tighter tolerance design (b).



(a)



(b)

Figure 6.6: Effects of different variation sources on signal delay with optimally buffered interconnect using worst-case tolerances (a) and tighter tolerance design (b).

of buffer delay variation increases (from 60 psec to over 80 psec) since more buffers must be inserted for a fixed interconnect length. From Fig. 6.6 (b), we see that tightening the poly CD variation tolerance results in a reduction of the total delay variation (e.g. from 95 psec to 55 psec at minimum pitch).

6.3 Impact of Variation on Crosstalk Noise

We now study the impact of interconnect variation on crosstalk noise for parallel wires in the global tier using 0.25 μm technology parameters. As discussed in Chapter 2, crosstalk noise depends on interconnect length, signal rise time, and driver size. We look at the effects of each of these parameters in addition to the effects of metal thickness and metal linewidth variation. The configuration used is equivalent to that shown in Fig. 1.14, and the interconnect is modeled as a distributed RC network. Hspice simulations are performed and the peak crosstalk noise is measured at the output of the victim (center) line while both neighbors switch in phase simultaneously. The interconnect and device parameters are given in Table 6.4.

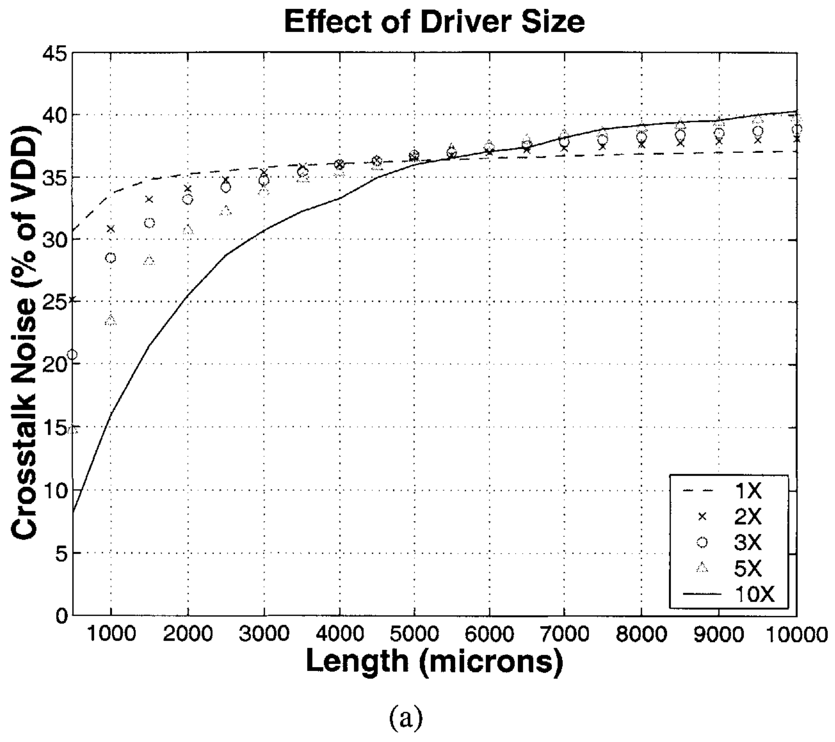
Fig. 6.7 shows the crosstalk noise as a function of interconnect length for different driver sizes and rise times. The crosstalk noise is plotted as a percentage of the supply voltage, V_{DD}. Crosstalk increases with increasing interconnect length, and begins to saturate at around 10 mm. We see that the effect of driver size is very important, especially for short wires. For a 500 μm wire, a minimum size driver results in crosstalk noise of over 30%, while a driver of 10X minimum size results in a peak crosstalk noise of less than 10% (see Fig. 6.7 (a)). It is also interesting to note that increasing the driver size results in larger crosstalk for lines greater than about 5 mm. Another important issue is the rise time of the signal. Fast rising signals cause more crosstalk noise than those with slower rise

times (see Fig. 6.7 (b)). For a 500 μm wire, the difference between a 100 and 300 psec rise time at the driver input results in almost a 10% difference in the crosstalk noise.

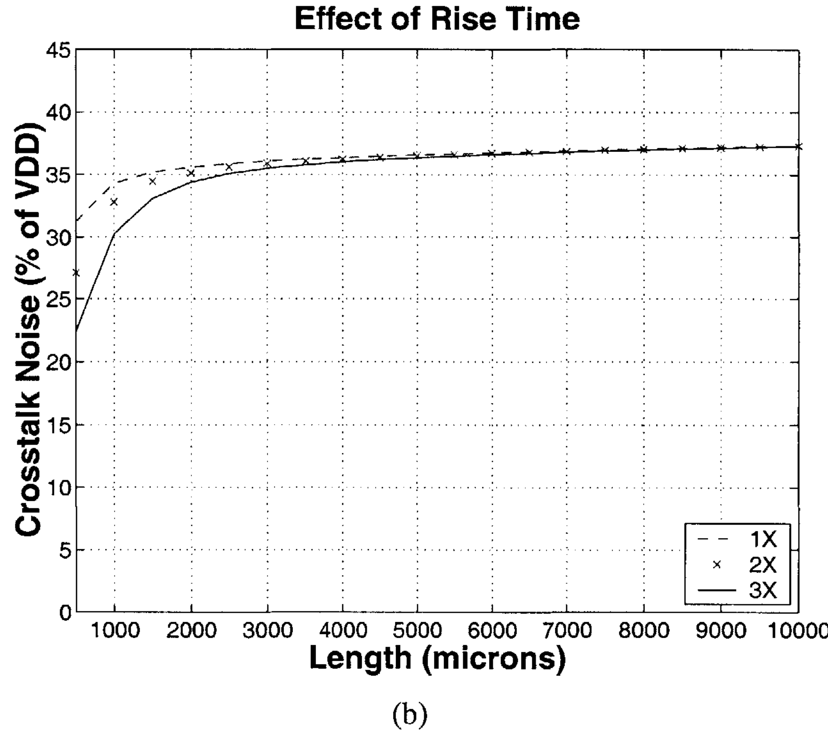
Table 6.4: Parameters for Crosstalk Noise Simulations

Parameter	Value	Units
V_{DD}	2.5	V
T, H	2.0	μm
Minimum LW, LS	0.75	μm
Minimum Rise Time	100	psec
Driver Gate Length L_{eff}	0.25	μm
NMOS Gate Width W_{gate} (Min Size Driver)	0.375	μm
Driver Resistance R_0 (Min Size Driver)	620	Ω
Transistor Load Capacitance C_0 (Min Size Driver)	8.2	fF
LW, LS Variation	0.075	μm
T Variation	0.20	μm

Fig. 6.8 shows the effect of pitch and metal thickness and linewidth variation on crosstalk noise for interconnects of varying length. From Fig. 6.8 (a) we see that increasing the pitch (constant linewidth and increasing linespace from 1X to 3X) results in a reduction in the crosstalk noise by about 10% for lines between 500 μm and 10 mm. Looking at Fig. 6.8 (b), we see that metal thickness variation results in a lower than nominal crosstalk noise and metal linewidth variation results in a greater crosstalk noise. This is expected since metal CMP variation causes a reduction in the metal thickness, resulting in less intra-layer coupling. On the other hand, an increase in the linewidth causes more coupling between adjacent wires and therefore an increase in the crosstalk noise. Comparing

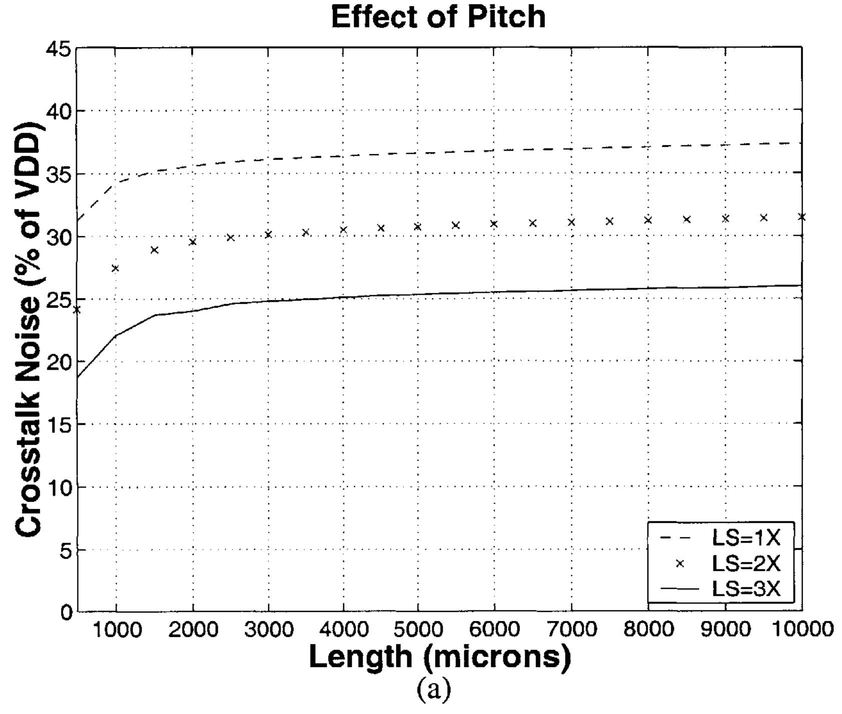


(a)

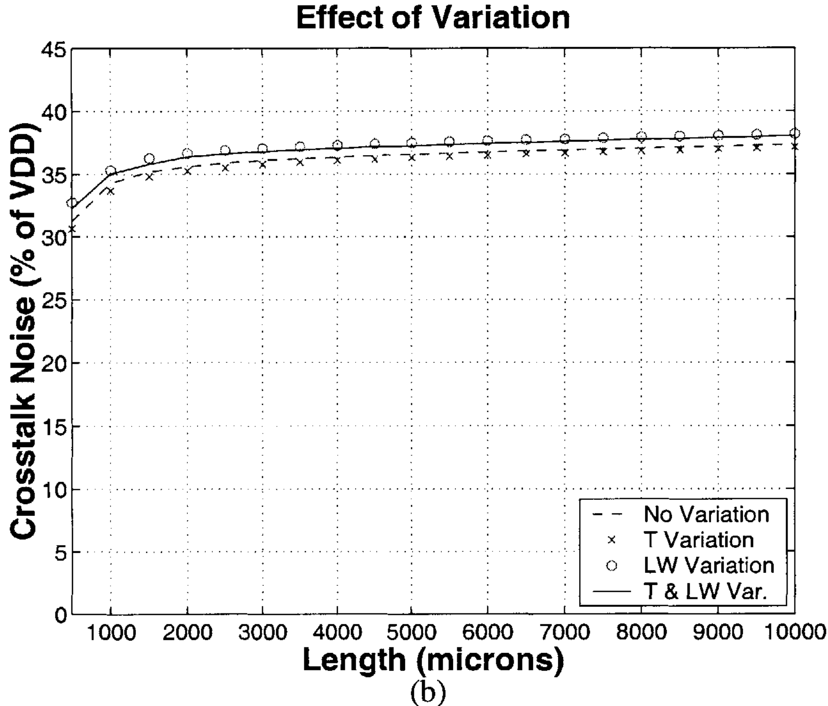


(b)

Figure 6.7: The effect of driver size (a) and rise time (b) on crosstalk noise vs. interconnect length.



(a)



(b)

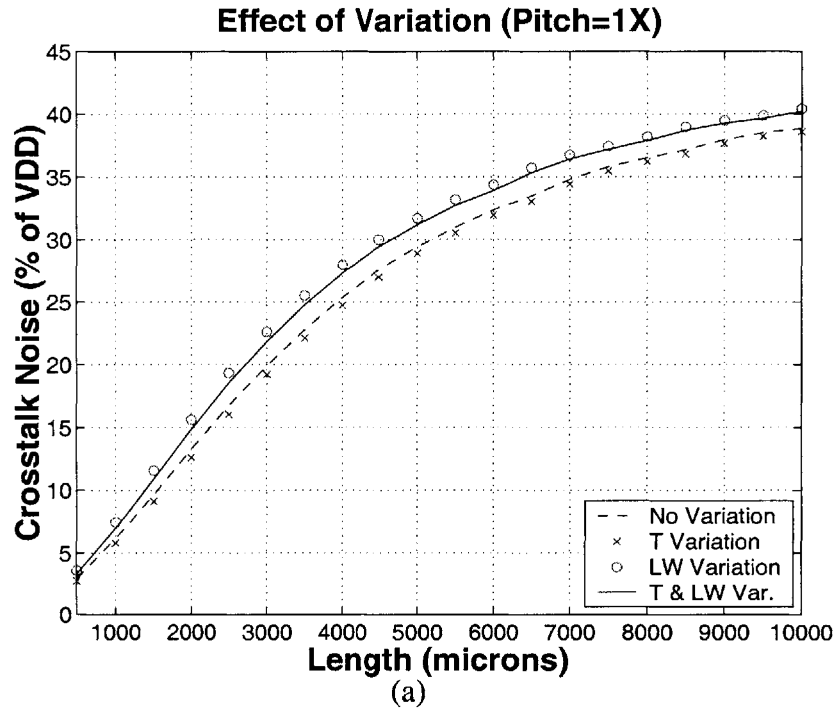
Figure 6.8: The effect of pitch (a) and effect of metal thickness and linewidth variation (b) with minimum size drivers and input rise time of 100 psec.

the effects of variation with those of driver size, rise time, and pitch, we find that variation does not seem to impact crosstalk noise as much.

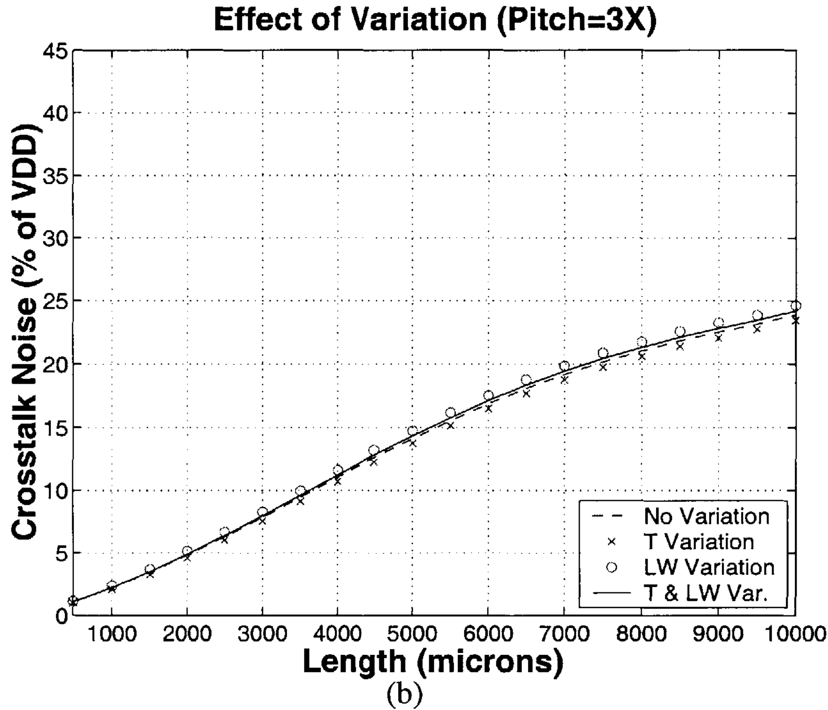
From Fig. 6.7 we have seen that increasing the driver size (up to interconnect lengths of about 5 mm) and rise time decreases the crosstalk noise. Since a good interconnect design minimizes the crosstalk noise, we also simulate the effects of variation for a 10X driver size and a 300 psec aggressor input rise time. The results are shown in Fig. 6.9. For an acceptable crosstalk noise criteria of 20% of V_{DD}, the maximum interconnect length is 3 mm without any variation at minimum pitch. With variation, it is reduced to about 2.7 mm. At 3X minimum pitch, the effect of variation is even less.

6.4 Variation Impact on Crosstalk With Optimal Interconnect

In the previous section, we studied the effects of variation on crosstalk noise as a function of different interconnect and device parameters such as interconnect length, driver size, and input rise time. Here, we consider signals that are optimized for minimum delay and determine the impact on crosstalk for global wires in the 250 nm technology generation. The interconnect length for these simulations is taken as the distance between two adjacent buffers for the case of optimally buffered interconnect, and it depends on the wiring pitch. Since there is only one segment between adjacent buffers, the number of buffers is equal to one. The optimal interconnect length and buffer size are given in Table 6.5 for different pitch.



(a)



(b)

Figure 6.9: Crosstalk noise vs. interconnect length for driver size of 10X and rise time of 300 psec with and without variation for minimum pitch (a) and 3X pitch (b).

Table 6.5: Interconnect Length and Buffer Size for Optimized Crosstalk Simulations

Pitch	Interconnect Length (μm)	Buffer Size
1X	1455	40.44
2X	1882	31.66
3X	2133	28.79

The crosstalk noise is simulated for a fixed input rise time of 100 psec. Fig. 6.10 shows the crosstalk as a percentage of the supply voltage for the case without any variation as well as with worst-case variation and tighter tolerance design. The results show that although crosstalk is a function of pitch, the effect of variation is not as significant as it is on signal delay. We see that with worst-case variation, crosstalk increases by no more than 2% of V_{DD}. If tighter tolerances are used, the increase is only 1% compared with the case without any variation. This implies that for the 250 nm generation, including the effects of variation (systematic or random) are not as important in crosstalk simulation as they are for signal delay and clock skew.

6.5 Summary

In this chapter, we have studied the effects of both systematic and random variation on clock skew, signal delay, and crosstalk noise. We have seen that in addition to modeling the systematic variation, it is also necessary to include the effects of random variation in circuit simulation. Our clock skew case study shows that skew due to random device variations accounts for up to 10% of the clock cycle in a 1 GHz design, whereas systematic variation effects account for about 3% of the clock cycle. When every picosecond counts, modeling the systematic CMP effects and tightening the CD variation tolerances can pro-

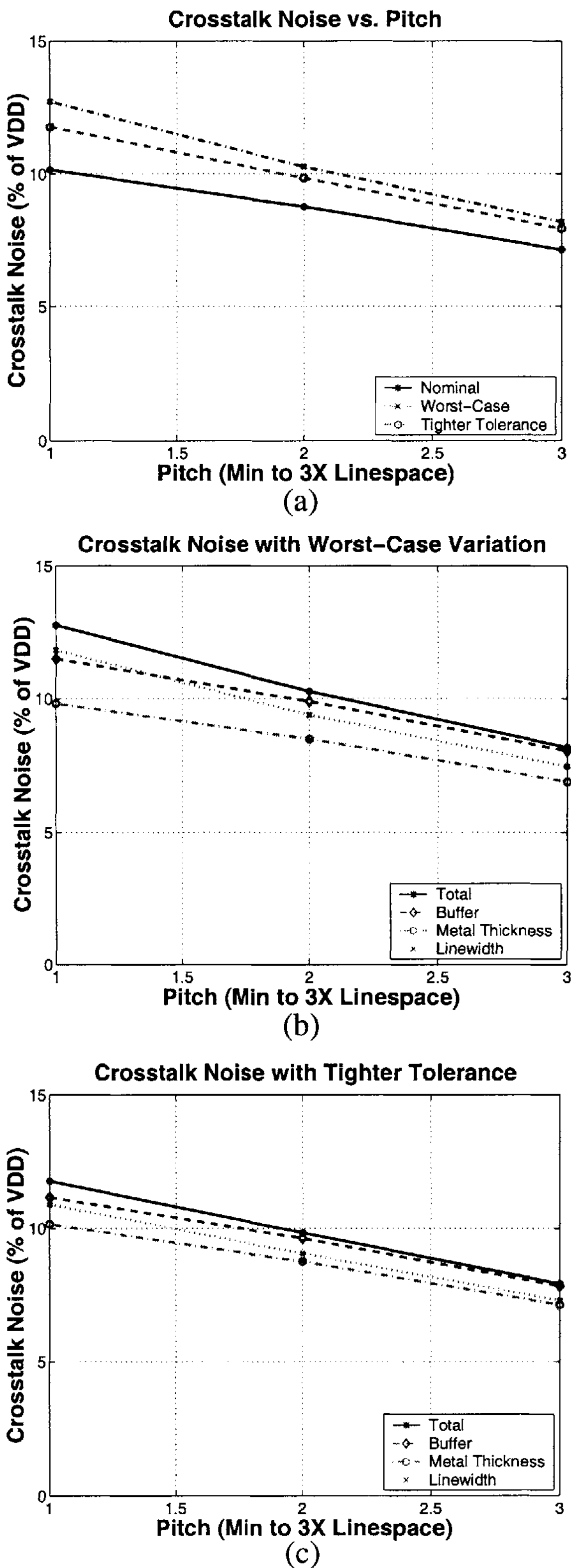


Figure 6.10: Crosstalk noise for no variation, worst-case variation, and tighter tolerance design (a). Effects of different variation sources on crosstalk with worst-case tolerances (b) and tighter design tolerances (c).

vide designers with an aggressive alternative to unnecessary overdesign. Rather than increasing buffer sizes to over-drive the clock signal (which use up extra chip area and increase the power dissipation), systematic variation models can help reduce the design uncertainty while meeting the design objectives. In our example of the buffered clock tree in this chapter, systematic variation modeling actually resulted in a lower skew than the case without any variation. This is because the effects of asymmetry combined with variation resulted in some cancellation. A worst-case approach, however, would always predict a higher skew than the case without any variation. Similarly, in the case of optimally buffered interconnect, we are able to increase the maximum interconnect length by up to 10% or more through tighter tolerance design. As technology scales, worst-case design approaches can impose unnecessary limitations to circuit design. In the next chapter, we scale our test circuits up to the 50 nm generation to determine the impact of variation with scaling.

Chapter 7

Technology Scaling Impact

This chapter highlights the increasing importance of systematic variation modeling with technology scaling [69]. The SIA Roadmap predicts that significant advances will need to be made in order for technology scaling to continue at its current pace and lists interconnect as one of the important areas where innovation will be required. Many recent publications describe the limits of technology scaling, and it appears that new interconnect modeling techniques are becoming necessary to achieve acceptable desired performance when variation is taken into account.

7.1 Technology Scaling

In [70, 71], several limiting factors to technology scaling and giga-scale integration (GSI) are described. These include theoretical and practical limits (fundamental, material, device, circuit, and system) with interconnect being the major circuit constraint [70]. It is shown that although scaling reduces device delay, interconnect delay continues to increase since chip size increases as more functionality is put onto a single die.

Given the approaching limits on technology, many works have projected the impact of technology scaling and have proposed methods to optimize interconnect design. One approach is to increase the vertical interconnect pitch along with the horizontal pitch [71]. Global interconnect is optimized by sizing the vertical and horizontal pitch and is optimized mainly for delay. Local interconnect is optimized for crosstalk and electromigration through the use of selective driver sizing. However, according to [72], innovations in the interconnect and ILD materials will be needed eventually since increasing the metal thickness/width aspect ratio above two does not reduce the delay very much.

In other works [73, 74], architectural interconnect design optimization is described. In [73], the complete wiring distribution (local, intermediate, global) based on Rent's Rule is first calculated. Then a methodology is proposed to optimize the architecture (minimize number of wiring levels and maximize interconnect density) based on clock frequency, power consumption, and chip area constraints [74]. In [75, 76], the impact of technology scaling is studied on global wires using the global wiring distribution technique given in [73]. Wire pitch sizing, repeater insertion, and the use of new materials is explored to optimize wiring delay and determine wiring requirements. Trade-offs are described, and [75] concludes that new low resistivity, low dielectric constant materials will be needed for the interconnect by the 0.10 μm generation.

In the following sections, we study the impact of scaling on the different performance metrics, but also incorporate variation effects in this scaling. First, the clock tree is scaled to the 50 nm generation. The impact of interconnect and device variation on clock skew is simulated. Next, the effects of scaling and variation are simulated for interconnect for technologies ranging from the 250 nm to 50 nm generations. A clock frequency constraint is placed on the path delay of optimally buffered interconnect, and the effect of variation on the maximum interconnect length is computed for interconnect of different pitch wired in the local, intermediate, and global tiers. We study the impact due to different "variation scaling scenarios" to determine the relative importance of variation modeling in future generations compared to today.

7.2 Technology Scaling Impact on Clock Skew

In this section, we consider once again the clock skew in the 1 GHz H-tree from the previous chapters. Here, the tree is scaled from the 180 nm technology to the 50 nm generation. The device and power supply variation scaling is performed according to the SIA

Roadmap and the scaling projections given in [64], as shown in Tables 7.1 and 7.2. Table 7.1 gives the nominal and 3σ percentage variation for L_{eff}, t_{ox}, V_t, and V_{DD}. Note that as technology scales to the 50 nm generation, all of the variation tolerances are not expected to scale at the same rate. Although the V_{DD} variation tolerance is expected to remain constant at 10%, the L_{eff} variation tolerance is expected to double from 20% to 40%. Table 7.2 provides the parameters for the scaled clock tree used in our clock skew simulation. The interconnect pitch is expected to shrink to 0.25X, while the chip side length is expected to double to 34 mm and the global clock frequency triples to 3 GHz.

Table 7.1: Scaled Device and Power Supply Parameters and 3σ Variations

Parameter	Nominal	Variation (%)	Nominal	Variation (%)
L_{eff} (nm)	180	20	50	40
t_{ox} (nm)	2.2	10	0.7	20
V_t (V)	0.40	10	0.25	15
V_{DD} (V)	1.8	10	0.9	10

Table 7.2: Scaled Clock Tree Parameters

Parameter	180 nm	50 nm
LW, LS	1X	0.25X
Aspect Ratio	1X	1.3X
Dielectric Constant	3.5	1.4
Chip Size	17 mm x 17 mm	34 mm x 34 mm
Global Clock Frequency	1 GHz	3 GHz

7.2.1 Clock Skew

Table 7.3 summarizes the clock skew results for the H-tree scaled to the 50 nm generation. The clock skew without any variation is reduced from 46 psec in the 180 nm generation to 22 psec in the 50 nm generation. Although the interconnect pitch shrinks and chip size increases as technology scales, the relative dielectric constant of the inter-layer dielectric is expected decrease from 3.5 to 1.4 and the metal resistivity is expected to decrease from 2.2 $\mu\Omega$-cm to 1.6 $\mu\Omega$-cm by the 50 nm generation (see Table 7.4). The decrease in both the dielectric constant and resistivity reduce the interconnect capacitance and resistance, and therefore also reduce delay and skew.

As technology scales, the importance of systematic variation modeling (especially poly CD) increases. A skew reduction of 25 psec is obtained in the 50 nm generation, which is comparable to the 28 psec reduction in the 250 nm generation, by modeling the metal thickness variation and tightening the poly CD tolerance. However, 25 psec corresponds to 7.5% of the clock cycle in the 50 nm generation, whereas 28 psec corresponds to only 2.8% of the clock cycle in the 180 nm generation. Given that the total clock skew budget is around 10-15% of the clock period, implementing systematic variation effects will prove to be quite useful in achieving an acceptable clock skew budget. Although increasing buffer sizes can lower the skew, it comes at the expense of additional power consumption. The trade-off between increasing buffer sizes to reduce skew while increasing power consumption will be a concern as technology scales since low power design is already an important issue.

Table 7.3: Technology Scaling Impact on Clock Skew

Variation Source	180 nm	50 nm	180 nm	50 nm
None	45.53	22.12	4.55	6.62
Cu CMP (systematic model)	36.85	14.83	3.69	4.44
Devices & V_{DD}	116.58	71.28	11.66	21.34
Devices & V_{DD} (tighter L_{eff} tolerance)	96.96	53.40	9.70	15.59
Total Skew (worst-case device & V_{DD} tolerances)	162.11	93.40	16.21	27.96
Total Skew (systematic model for metal thickness & tighter L_{eff} tolerance)	133.81	68.23	13.38	20.43
Skew Reduction (systematic models)	**28.30**	**25.17**	**2.83**	**7.53**
Units	**psec**		**% of clock period**	

7.3 Interconnect and Device Variation Scaling

In Table 7.4, selected interconnect parameters taken from the SIA Roadmap are listed for technologies ranging from the 250 nm to 50 nm generation. Some of the numbers given in the roadmap are for local interconnect, and we scale these numbers for intermediate and global interconnect, assuming a three-tier wiring configuration as is common in most high performance designs. The interconnect pitch is equal to *LW+LS* and the aspect ratio is given as *T/LW.* The metal and ILD thickness aspect ratios for the three-tier system are based on the roadmap. Aspect ratios of 2, 2.5, and 2.67 are used for the local, intermediate, and global wiring layers. For our study, we let linewidth equal to linespace and metal thickness equal to ILD thickness.

Table 7.4: Interconnect Technology Parameters

Parameter/ Technology	**250nm**	**180nm**	**130nm**	**100nm**	**50nm**
Local Interconnect Pitch (nm)	500	360	260	200	100
Intermediate Pitch (nm)	1000	720	520	400	200
Global Pitch (nm)	1500	1080	780	600	300
Dielectric Constant	4.0	3.5	2.7	2.0	1.4
Metal Resistivity ($\mu\Omega$-cm)	2.2	2.2	2.2	2.2	1.6
Global Clock Frequency (MHz)	1000	1200	1600	2000	3000

Fig. 7.1 shows the interconnect resistance, capacitance, and RC per unit length as a function of the technology generation for local, intermediate, and global interconnect. As technology scales, the total RC increases, indicating that interconnect contributes a larger fraction of the total path delay. In addition, the global clock frequency is expected to triple and the local clock frequency will be five times that of today by the 50 nm generation.

Given an interconnect scaling scenario and the required clock frequency f_c, the maximum interconnect length allowed L_{max} and the optimal number of buffers *bufopt* is calculated. To put this in perspective, we plot the maximum length that we can route in one clock cycle in Fig. 7.2 for local, intermediate, and global interconnect. The maximum required interconnect length is $2L_{chip}$, where L_{chip} is the chip side length. If $L_{max} \geq 2L_{chip}$, it is possible to route between two points on the chip in one clock cycle

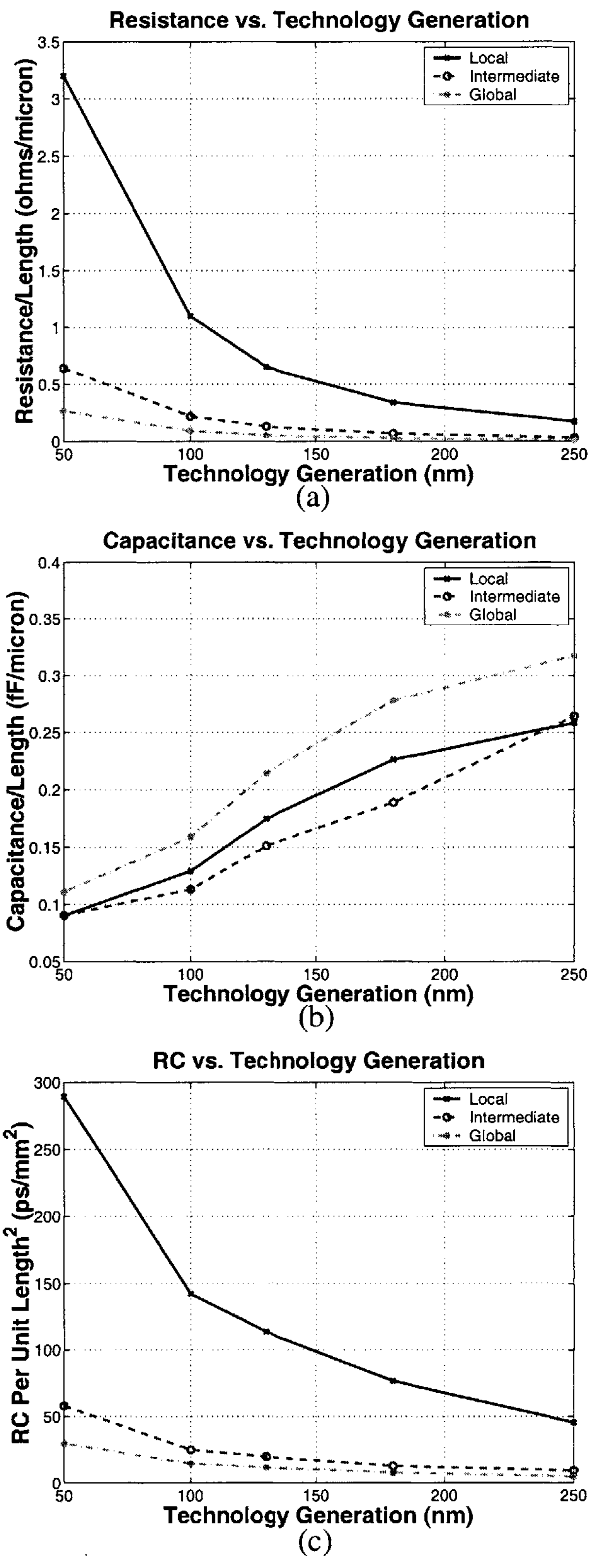


Figure 7.1: The interconnect resistance (a), capacitance (b), and RC (c) per unit length for minimum pitch wires in the local, intermediate, and global tiers.

(assuming ideal wiring conditions and no other stages of logic in between). We denote the transition point as the technology generation where L_{max} is equal to $2L_{chip}$. Generally, global and intermediate wiring layers are used to route over long distances on the chip. At minimum pitch, the transition point for the intermediate wiring tier occurs between the 180 nm and 130 nm generations. For global wires, the transition point is between the 130 nm and 100 nm generations. Increasing the wiring pitch (usually linespace) decreases the intra-layer coupling capacitance and thus reduces wiring delay. Note that even for global wires at 3X minimum pitch, $L_{max}<2L_{chip}$ by the 50 nm generation.

7.3.1 Interconnect Variation Scaling Scenarios

Several variation scaling scenarios are possible and it is not certain that the percentage tolerance on each variation source will scale with technology. There are several possible scenarios for variation scaling, the first being that a constant absolute tolerance is maintained. However, this is very unlikely since there will be improvements in processing technology. These improvements include pushing technology further and closer to its limits and better manufacturing equipment and process control. We start with the 250 nm generation and consider three variation scaling scenarios:

(i) Constant percent scaling,

(ii) Variable percent scaling (case 1),

(iii) Variable percent scaling (case 2).

Scenario (i) is the ideal case, where the same percentage design tolerances are maintained with scaling. Scenarios (ii) and (iii) represent different scaling rates. In both cases 1 and 2, we assume that the absolute tolerance is decreasing but the percentage tolerance is increasing. For case 1, we assume that the percentage tolerances double by the 50 nm, and for case 2 we assume that the tolerances triple by the 50 nm generation. We use a linear

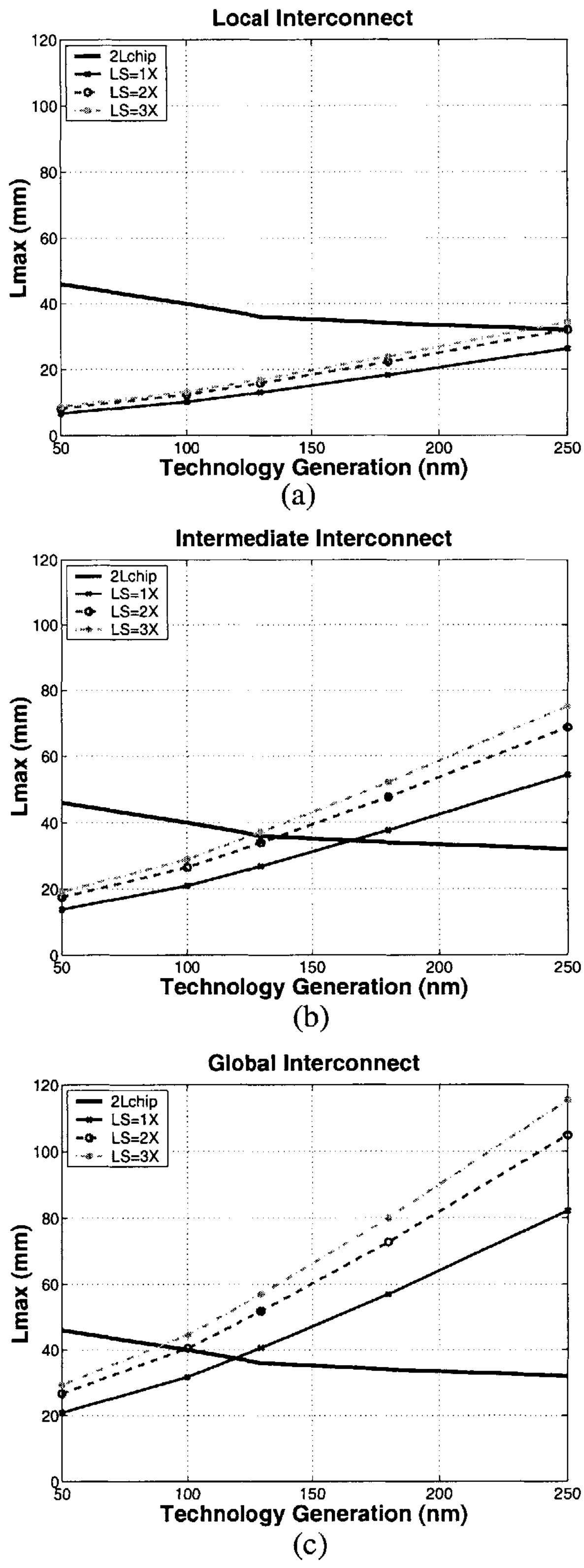


Figure 7.2: Maximum interconnect length (optimally buffered to satisfy clock frequency constraint) vs. technology generation for local (a), intermediate (b), and global interconnect (c).

scaling approach to determine the worst-case variation tolerances for the intermediate technology generations. The projected variations for the three scenarios are given in Table 7.5 as a percentage of the minimum linewidth and nominal metal thickness.

Table 7.5: 3σ Variation Tolerances for 250 nm and 50 nm Generations

Tier	Technology (nm)	Scaling Scenario	Variation in Percent (LW, LS, Metal Thickness)
Local	250	Constant Percent	20
Intermediate	250	Variable Percent (2X)	15
Global	250	Variable Percent (3X)	10
Local	50	Constant Percent	20
Local	50	Variable Percent (2X)	40
Local	50	Variable Percent (3X)	60
Intermediate	50	Constant Percent	15
Intermediate	50	Variable Percent (2X)	30
Intermediate	50	Variable Percent (3X)	45
Global	50	Constant Percent	10
Global	50	Variable Percent (2X)	20
Global	50	Variable Percent (3X)	30

7.3.2 Device Variation Scaling

In addition to the interconnect variation, device variation must also be considered for a comprehensive perspective on the overall effects of variation. The device parameters are

scaled according to the SIA roadmap. The variation tolerances are scaled for scenarios (ii) and (iii) according to the projections given in [64] (see Table 7.1). Note that just as for the interconnect, the device variation is also not expected to scale with technology. In particular, the L_{eff} and t_{ox} tolerances are projected to double from 20% to 40% and 10% to 20%, respectively.

7.4 Variation Impact With Technology Scaling

We study the impact of variation on the maximum allowed interconnect length L_{max}, given the clock frequency constraint $f \geq f_c$ and optimally buffered interconnect. We look at the variation in the interconnect linewidth and the metal thickness. Assuming that a damascene process with metal (copper) CMP is used, there should not be any systematic variation in the ILD thickness. Metal linewidth (*LW*) and linespace (*LS*) variation must also be considered in addition to the device variation. Using the scaling scenarios given in Tables 7.1 and 7.5, we calculate the effect of variation on L_{max}.

7.4.1 Effect of Scaling Scenarios on L_{max}

Fig. 7.3 shows the effect of all the variation sources combined on L_{max}. The impact of variation on the L_{max} degradation is about 20-25% in the 250 nm generation depending on the wiring tier. As expected, we see that the impact on L_{max} stays approximately the same (or even decreases slightly) if the percentage worst-case tolerances remain constant for both devices and interconnect. For the cases of increasing variation tolerances, the interconnect length degradation increases. By the 50 nm generation, the L_{max} degradation is over 30% for global interconnect and 40% for local interconnect for scaling scenario (ii), where the interconnect variation tolerances double. The greatest sensitivity to variation is for *local* interconnect. The reason for this is that local interconnect is sensitive to both device and interconnect variation. Fine pitch (local) lines have a greater RC time constant

as well as a greater sensitivity since the interconnect variation tolerance is lower for global interconnect within a given technology (even though absolute tolerance may be greater).

7.4.2 Worst-Case Models vs. Tighter Tolerance Design

Although constant percent scaling may not be possible (at least for all variation sources), we can take advantage of our approach of tighter tolerance design, where systematic variation models are used for the CMP metal thickness variation and the poly CD variation tolerance is reduced by 50%. Fig. 7.4 shows the effect of variation on L_{max} for local and global interconnect using both the worst-case and tighter tolerance design approaches based on scaling scenario (iii). The results are plotted for wires of minimum pitch up to 3X pitch (with constant linewidth and variable linespace as before). We find that a significant gain in the interconnect length can be obtained by reducing design uncertainty. Also, the gain increases as technology scales using our tighter tolerance method, making the approach even more useful. About a 15% gain is possible for global interconnect and 20% for local interconnect in the 250 nm generation compared to less than 10% in the 50 nm generation with the tighter tolerance design method. Note that as pitch increases the L_{max} degradation also increases, even though the nominal value of L_{max} is larger at 3X pitch.

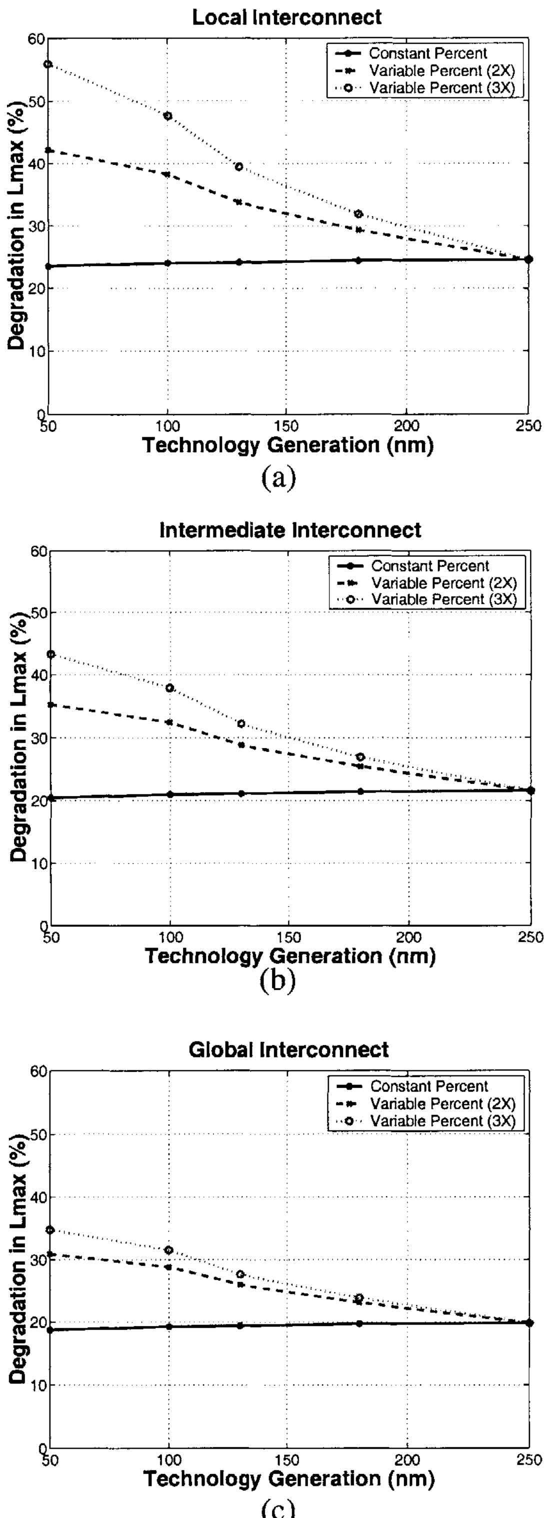
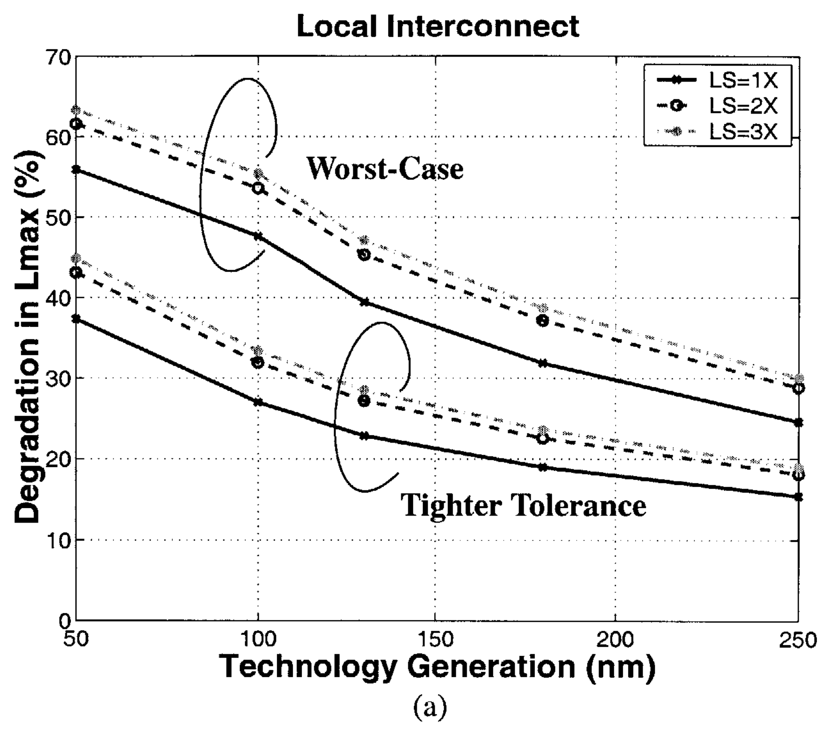
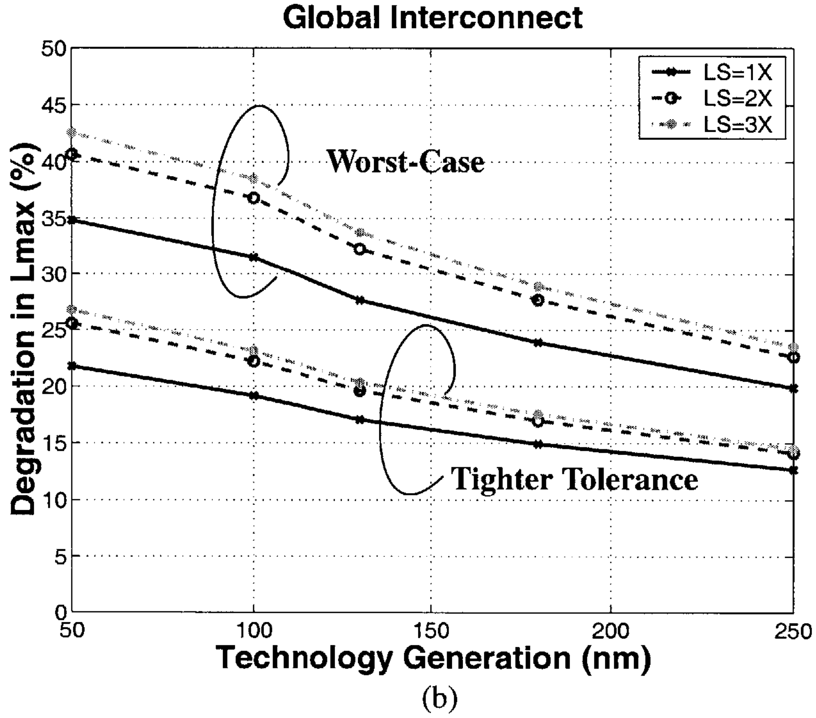


Figure 7.3: The degradation in L_{max} for optimally buffered local (a), intermediate (b), and global (c) interconnect for different variation scaling scenarios using worst-case tolerances.



(a)



(b)

Figure 7.4: The degradation in L_{max} for optimally buffered local (a) and global (b) interconnect using worst-case variation tolerances and tighter tolerance design.

7.4.3 Scaling Scenario Comparisons: Effect on L_{max}

Although variable scaling is probably most likely, at what rate the worst-case interconnect (and device) design tolerances increase will depend on the rate of process technology improvements. Let us compare the effects of different variation sources on L_{max} for the intermediate scaling scenario (scenario (ii)). The effects of each variation component (linewidth, metal thickness, and device variations) are shown in Fig. 7.5 for minimum pitch wires in the local and global tiers.

For global interconnect, linewidth and metal thickness variation have the smallest effect on delay, resulting in an L_{max} degradation of about 1% in the 250 nm generation. Even as technology scales to the 50 nm generation, the degradation in L_{max} increases to only about 2%. However, *buffer* delay variation is a significant concern. Device variations result in an L_{max} degradation of 12% in the 250 nm and increase to 18% by the 50 nm generation. For local interconnect, technology scaling places a larger emphasis on metal thickness variation, where the L_{max} degradation increases from 3% in the 250 nm generation to 15% by the 50 nm generation. Surprisingly, linewidth variation is not as large a concern since the L_{max} degradation resulting from linewidth variation is only 4% even in the 50 nm generation. One explanation is that linewidth variation affects both intra- and inter-layer capacitance, offsetting the (larger) variation in interconnect resistance. With metal thickness variation, the lateral coupling capacitance is affected most, not offsetting the increase in resistance as much. However, the buffer delay variation is still the most important, resulting in close to a 20% L_{max} degradation in the 50 nm generation.

There are two important points to note from the results shown in Figs. 7.4 and 7.5. First, as technology scales, the degradation in L_{max} increases due to systematic and random variation. Second, the L_{max} degradation computed using worst-case limits results in a

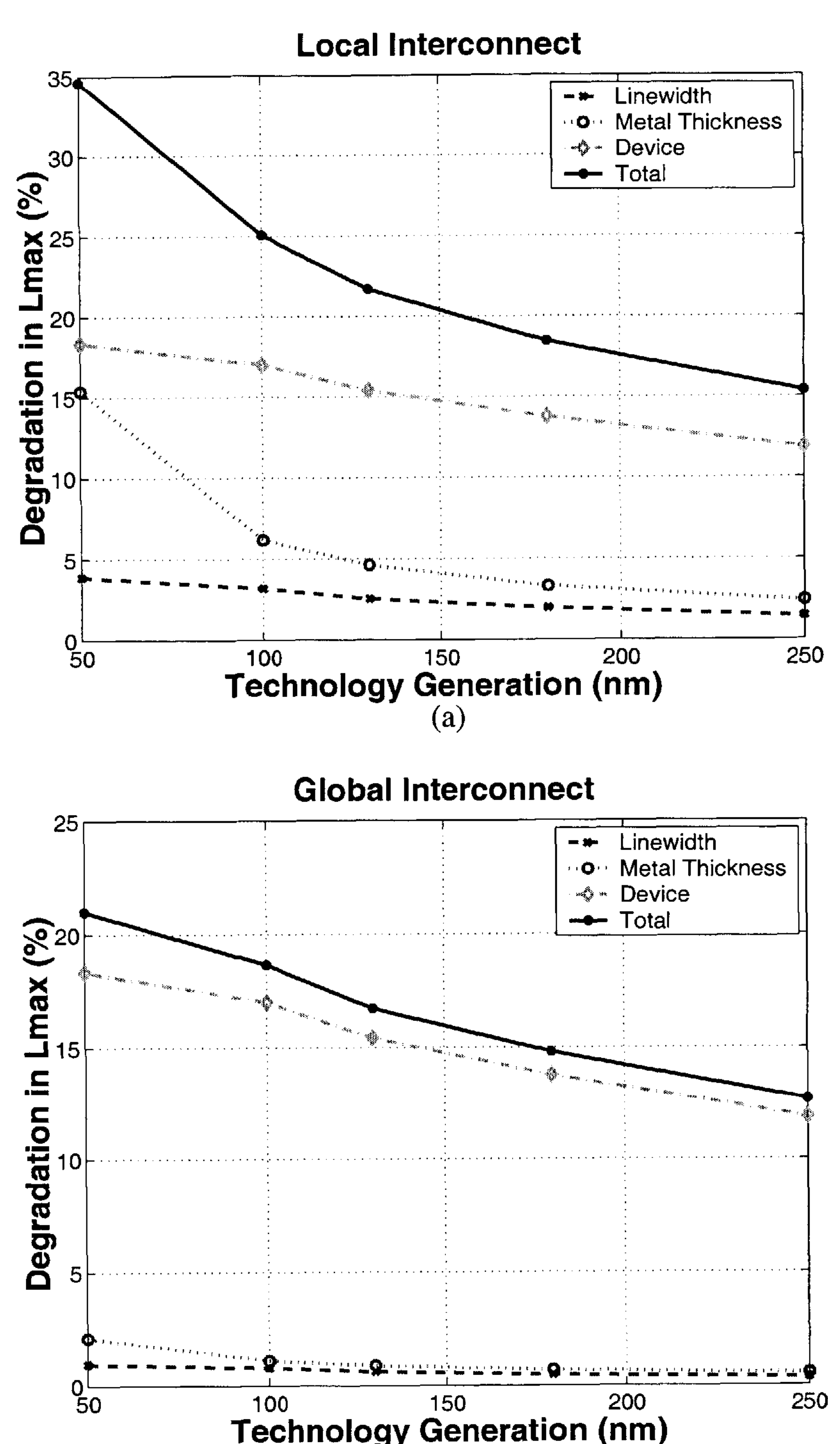


Figure 7.5: The degradation in L_{max} due to metal thickness, linewidth, device, and total variation for local (a) and global (b) interconnect at minimum pitch based on variable percent scaling scenario (ii).

greater over-estimate with technology scaling. In other words, continuing to use worst-case limits instead of systematic models in the 50 nm generation forces us to throw away a larger fraction of the clock cycle.

It is clear that to maintain the same level of variation tolerance on delay, a constant percent scaling of the variation tolerances is necessary. If one sets a threshold on the maximum allowable tolerance or safety margin on delay that will be attributed to random effects, one can determine the technology generation at which systematic variation models should be used without "throwing away" valuable design margin. The value of L_{max} calculated in Fig. 7.2 is the maximum length allowable for routing before a register (or pipeline stage) must be added assuming that there are no other logic gates in the path. Realistically, there will be logic in between and the actual value of L_{max} will be smaller than that obtained from Fig. 7.2. Two conditions must be satisfied under which systematic models are required (although they may be required earlier):

(i) $L_{max} < 2L_{chip}$

(ii) Degradation in $L_{max} > tol$ (tolerance)

With the tolerance set at around 20%, systematic variation modeling is already an important issue, and its importance will increase as technology scales.

7.5 Technology Scaling Impact on Crosstalk Noise

We have seen in Chapter 6 that for optimally buffered global interconnect in the 250 nm generation, the impact of worst-case neighbor switching results in crosstalk noise that is less than 15% of V_{DD}. We have also seen that variation has a small impact on crosstalk. Here, we consider the effects of technology scaling. In Fig. 7.6, the crosstalk noise is plotted vs. technology generations between 250 nm and 50 nm for different pitch. It is interesting to note that the crosstalk noise actually *decreases* as technology scales. One might

expect that the crosstalk noise will increase since interconnect pitch shrinks and there is more coupling between adjacent neighbors. However, if the interconnect is optimally buffered, the maximum wire length between buffers also decreases. Since crosstalk noise has a strong dependence on interconnect length, shorter lengths produce lower crosstalk. Our results show that crosstalk is reduced from 11% to 3% of V_{DD} in scaling from the 250 nm to 50 nm generation. The impact is not very large when variation is taken into account (see Fig. 7.7). Using the worst-case tolerances vs. tighter tolerance design method results in a fairly constant gain of about 1% of V_{DD} for all five generations. This difference is usually not significant but may be important in low swing circuits.

7.6 Summary

We have seen that it is important to model the effects of systematic variation in both interconnect and devices as technology scales. Considering the clock skew case study, we find that the systematic variation accounts for about 25 psec of skew in the 50 nm generation. While this value is comparable to the skew simulated with 250 nm generation parameters, the across chip global clock frequency is expected to triple by the 50 nm generation. This increases the relative importance of skew since 25 psec corresponds to over 7% of the clock cycle in the 50 nm generation. Considering the impact of variation in optimally buffered interconnect, we find that if variation tolerances do not scale at the same rate as the technology dimensions, using tighter tolerance design reduces the amount of overdesign that is required.

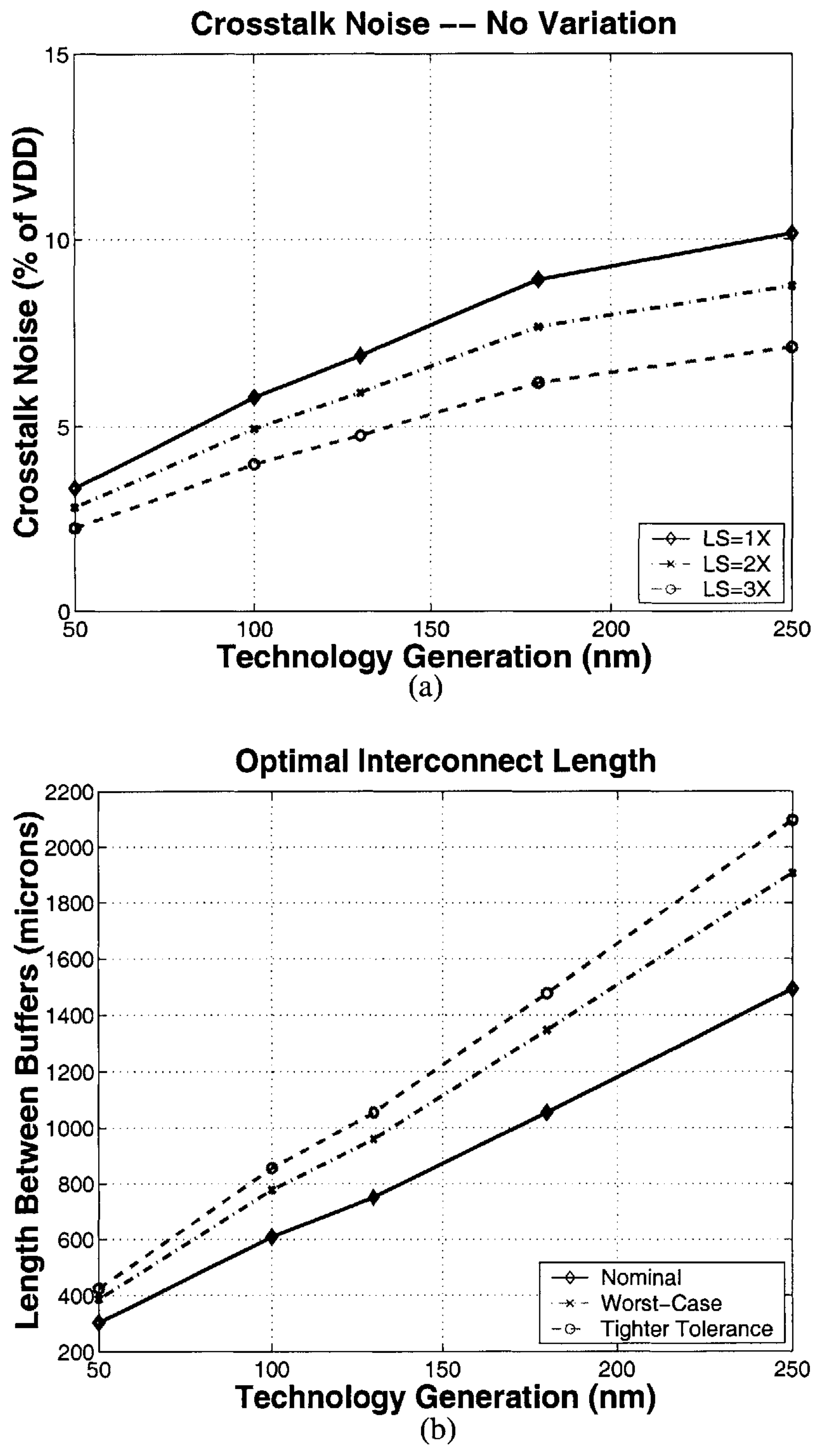


Figure 7.6: Crosstalk noise vs. technology generation for variable pitch with optimally buffered global interconnect (a). The optimal interconnect length between adjacent buffers varies as a function of the clock frequency constraint (b).

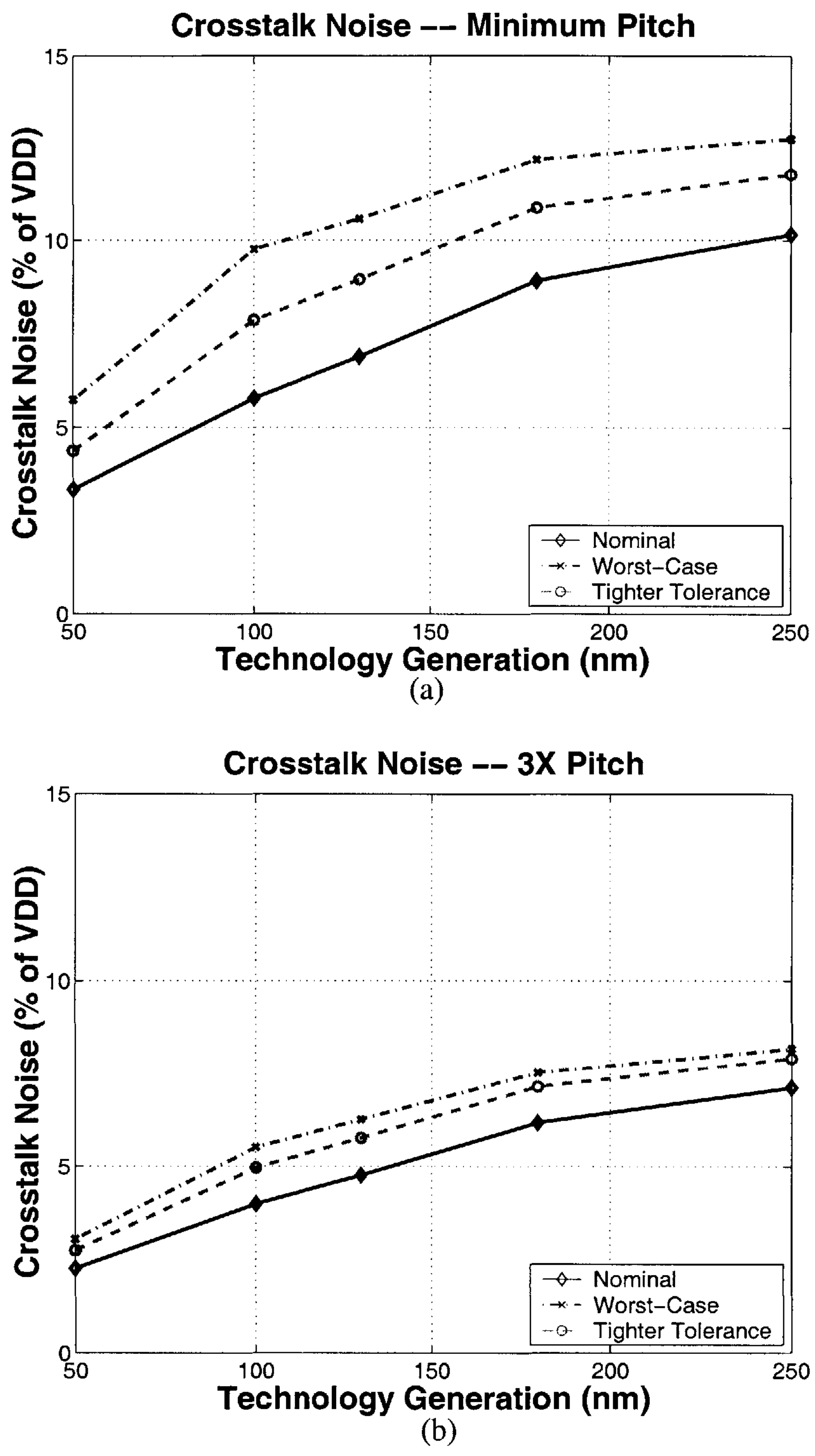


Figure 7.7: Crosstalk noise vs. technology generation for minimum (a) and 3X minimum (b) pitch with no variation, worst-case variation tolerances, and tighter tolerance design.

Chapter 8

Conclusion

8.1 Summary

In this thesis we have developed a new methodology to study the impact of variation on circuit performance. We start by separating the variation sources into systematic and random components. Variation impact for random components is simulated using conventional analysis techniques such as Monte-Carlo methods. The systematic components are modeled based on layout geometry and neighbor surroundings. These spatial characteristics strongly impact the interconnect variation and alter the metal linewidth, metal thickness, or ILD thickness. In addition, certain device variations such as the poly CD also have a systematic variation component. Spatial effects such as the linewidth, linespace, and effective pattern density need to be included to model the variation.

Our methodology uses existing CAD tools for circuit extraction and performance simulation, given a layout and technology information. During the extraction of nominal electrical parameters (e.g. interconnect resistance and capacitance, device gate length), information about the within-die coordinates (x, y) and neighbors within a specified distance away from the net or device of interest is also kept. This information allows us to perturb the netlist after it has been generated during the extraction process. The netlist is modified using our variation analysis tool, which is a collection of scripts that interface with the design tools. The advantage of our methodology is that the nominal electrical parameters only need to be extracted once. To include the effects of additional variation sources or different process conditions, the variation analysis tool needs to be run rather than doing a full re-extraction.

Once the methodology is developed and implemented, we study the effects of systematic process variation on industrial designs. These include global path delay variation and clock skew in 1 GHz microprocessors designed in the 0.25 μm and 0.18 μm technologies using aluminum and copper interconnect, respectively. We also study the relative impact of device variations in the clock tree, when intermediate buffers are inserted to drive the clock signal. Additionally, we simulate the impact of variation in optimally buffered interconnect to determine the impact of device vs. interconnect and random vs. systematic variations. Finally, the impact of variation is simulated to study the effects of technology scaling based on the SIA roadmap projections.

8.2 Conclusion

Our results show that systematic variation modeling can enable more aggressive design. This is clearly evident from our case studies. We have seen in the 1 GHz microprocessor study of ILD thickness variation that the design margin can be tightened if ILD thickness variation models are used instead of the worst-case limits. The simulation results show that up to a 48 psec tighter design margin can be obtained if a pattern density dependent ILD thickness variation model is used in circuit simulation. For the case of the buffered H-tree, we find that systematic variation modeling actually predicts a lower skew than the case without any variation because we are able to take advantage of the asymmetry in the circuit.

The technology scaling studies show that the importance of systematic variation modeling will increase in future generations. Our study of the clock tree shows that the reduction in the estimated clock skew that can be obtained with tighter tolerance design is over 25 psec in the 250 nm generation. This corresponds to about 2.5% of the clock cycle. Although a similar reduction in skew can be obtained in the 50 nm generation, it corre-

sponds to a much larger percentage (7.5%) of the clock cycle. This increased importance is also shown in the optimally buffered interconnect case study. We find that the difference in degradation in the maximum interconnect length of global wires increases from 7% to 15%, and 10% to 20% for local wires compared to the tighter tolerance design method.

Finally, an important result of our studies is that random device variations still need to be considered in interconnect dominant circuits. We find that random variations account for over 75% of the total variation in the clock skew and interconnect delay simulations. These variations are mainly due to the device parameters and the power supply variation. This implies that conventional approaches utilizing Monte-Carlo analysis techniques still need to be used. However, systematic variation modeling allows for more aggressive design when high performance is required. Tightening the margins can enable designers to focus their efforts on meeting or exceeding performance specifications.

8.3 Future Work

We have focused mainly on the impact of systematic variation models for the interconnect in this thesis. Specifically, we have considered the effects of CMP variation in substantial detail. An extension to our current work is to develop systematic variation models for the linewidth. This is an important area of research, especially since fine pitch interconnect is more sensitive to intra-layer coupling. Also, we have seen that poly CD variation is an important source of skew and delay variation. Another good extension to our methodology is to include the effects of inductance in circuit performance simulation. Since signal rise times are becoming faster and longer interconnect is being used, these effects have already started to become a concern and will be even more so in future technology generations. In addition, designing test circuits to compare the circuit performance

simulations with experimental observations will provide a method for calibrating the process variation models as well as provide more insight into other variation effects.

Finally, our technique of systematic variation analysis can be applied to study the impact on emerging interconnect technologies. These include variation issues in optical (opto-electronic) circuits for clock distribution and three-dimensional integrated circuits. Methods for variation modeling and analysis will continue to be an important area of research in future technologies.

Bibliography

[1] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 1999.

[2] H. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, 1990.

[3] J. Rabaey, *Digital Integrated Circuits: A Design Perspective*, Prentice Hall, 1996.

[4] D. Boning and S. Nassif, "Models of Process Variations in Device and Interconnect," in *Design of High Performance Microprocessor Circuits*, Editors: A. Chandrakasan, W. Bowhill, F. Fox, IEEE Press, 2000.

[5] C. Michael and M. Ismail, *Statistical Modeling for Computer-Aided Design of MOS VLSI Circuits,* Kluwer Academic Press, Boston, 1993.

[6] C. Chow, "Projection of Circuit Performance Distributions by Multivariate Statistics," *IEEE Trans. on Semi. Manuf.*, Vol. 2, No. 2, pp. 60-65, May 1989.

[7] O.S. Nakagawa, N. Chang, S. Lin, and D. Sylvester, "Circuit Impact and Skew-Corner Analysis of Stochastic Process Variation in Global interconnect," *Proc. International Interconnect Technology Conference*, pp. 230-232, 1999.

[8] D. Sylvester, O. S. Nakagawa, and C. Hu, "Modeling the Impact of Back-End Process Variation on Circuit Performance," *Proc. International Symposium on VLSI Technology, Systems, and Applications*, pp. 58-61, 1999.

[9] N. Chang, V. Kanevsky, O. S. Nakagawa, K. Rahmat, and S.-Y. Oh, "Fast Generation of Statistically-Based Worst-Case Modeling of On-Chip Interconnect," *Proc. IEEE International Conference on Computer Design: VLSI in Computers and Processors*, pp. 720-725, 1997.

[10] M. Orshansky, J. Chen, and C. Hu, "A Statistical Performance Simulation Methodology for VLSI Circuits," *Proc. DAC*, pp. 402-407, June 1998.

[11] M. Orshansky, C. Spanos, C. Hu, "Circuit Performance Variability Decomposition," *Proc. International Workshop on Statistical Metrology*, pp. 10-13, June 1999.

[12] Z. Lin, C. Spanos, L. Milor, and Y.-T. Lin, "Circuit Sensitivity to Interconnect Variation," *IEEE Trans. on Semi. Manuf.*, Vol. 11, No. 4, pp. 557-568, Nov. 1998.

[13] D. Boning, T. Maung, J. Chung, K.-J. Chang, S.-Y. Oh, and D. Bartelink, "Statistical Metrology of Interlevel Dielectric Thickness Variation," *Proc. SPIE Symposium on Microelectronics Manufacturing,* Vol. 2334, pp. 316-327, Austin, TX, Oct. 1994.

[14] E. Chang, B. Stine, T. Maung, R. Divecha, D. Boning, J. Chung, K. Chang, G. Ray, D. Bradbury, S. Oh, and D. Bartelink, "Using a Statistical Metrology Framework to Identify Random and Systematic Sources of Intra-Die ILD Thickness Variation for CMP Processes," *Proc. IEDM*, pp. 499-502, Wash. D.C., Dec. 1995.

[15] D. Boning and J. Chung, "Statistical Metrology: Understanding Spatial Variation in Semiconductor Manufacturing," *Proc. SPIE Symposium on Microelectronic Manufacturing: Microelectronic Manufacturing Yield, Reliability, and Failure Analysis II*, Oct. 1996.

[16] D. Bartelink, "Statistical Metrology: At the Root of Manufacturing Control," *Journal on Vacuum Science and Technology B*, Vol. 12, No. 4, pp. 2785-2794, Jul./Aug. 1994.

[17] B. Stine, D. Boning, and J. Chung, "Analysis and Decomposition of Spatial Variation in Integrated Circuit Processes and Devices," *IEEE Trans. on Semi. Manuf.*, Vol. 10, No. 1, pp. 24-41, Feb. 1997.

[18] J. Steigerwald, S. Murarka, and R. Gutman, *Chemical Mechanical Planarization of Microelectronic Materials*, John Wiley & Sons, 1997.

[19] R. Divecha, B. Stine, E. Chang, D. Ouma, D. Boning, J. Chung, O. S. Nakagawa, S.-Y. Oh, S. Prasad, W. Loh, and A. Kapoor, "Assessing and Characterizing Inter- and Intra-die Variation Using a Statistical Metrology Framework: A CMP Case Study," *Proc. First International Workshop on Statistical Metrology*, pp. 9-12, Honolulu, HI, June 1996.

[20] B. Stine, D. Ouma, R. Divecha, D. Boning, J. Chung, D. Hetherington, I. Ali, G. Shinn, J. Clark, O. S. Nakagawa, and S.-Y. Oh, "A Closed-Form Analytic Model for ILD Thickness Variation in CMP Processes," *Proc. Chemical Mechanical Polish for ULSI Multilevel Interconnection Conference (CMP-MIC)*, pp. 266-273, Santa Clara, CA, Feb. 1997.

[21] Raphael User's Guide, Technology Modeling Associates, 1997.

[22] K. Nabors and J. White, "Fastcap: A Multipole Accelerated 3-D Capacitance Extraction Program," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 10, No. 11, pp. 1447-1459, Nov. 1991.

[23] T. Sakurai and K. Tamaru, "Simple Formulas for Two- and Three-Dimensional Capacitances," *IEEE Trans. on Electron Devices*, Vol. ED-30, No. 2, pp. 183-185, Feb. 1983.

[24] T. Sakurai, "Closed-Form Expressions for Interconnection Delay, Coupling, and Crosstalk in VLSI's", *IEEE Trans. on Electron Devices*, Vol. 40, No. 1, pp. 118-124, Jan. 1993.

[25] J.-H. Chern, J. Huang, L. Arledge, P.-C. Li, and P. Yang, "Multilevel Metal Capacitance Models for CAD Design Synthesis Systems," *IEEE Electron Device Letters*, Vol. 13, No. 1, pp. 32-34, Jan. 1992.

[26] D. Cheng, *Fundamentals of Engineering Electromagnetics*, Addison-Wesley, 1993.

[27] J. Davis and J. Meindl, "Compact Distributed RLC Interconnect Models-Part II: Coupled Line Transient Expressions and Peak Crosstalk in Multilevel Networks," *IEEE Trans. on Electron Devices*, Vol. 47, No. 11, pp. 2078 -2087, Nov. 2000.

[28] K. Tang and E. Friedman, "Peak Crosstalk Noise Estimation in CMOS VLSI Circuits," *Proc. International Conference on Electronics, Circuits, and Systems*, Vol. 3, pp. 1539-1542, 1999.

[29] O. S. Nakagawa, D. Sylvester, J. McBride, and S.-Y. Oh, "On-Chip Cross Talk Noise Model for Deep-Submicrometer ULSI Interconnect," *The Hewlett-Packard Journal*, pp. 39-45, May 1998.

[30] A. Devgan, "Efficient Coupled Noise Estimation for On-Chip Interconnects," *Proc. International Conference on Computer-Aided Design*, pp. 147-151, 1997.

[31] M. Kamon, M. Ttsuk, and J. White, "Fasthenry: A Multipole Accelerated 3-D Inductance Extraction Program," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 42, No. 9, Part 1-2, pp. 1750-1758, Sept. 1994.

[32] L. He, N. Chang, S. Lin, and O. S. Nakagawa, "An Efficient Inductance Modeling for On-Chip Interconnects," *Proc. CICC*, pp. 457-460, 1999.

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

[34] T. Tugbawa, T. Park, D. Boning, T. Pan, P. Li, S. Hymes, T. Brown, and L. Camilletti, "A Mathematical Model of Pattern Dependencies in Copper CMP Processes," *Proc. CMP Symposium, Electrochemical Society Meeting*, Honolulu, HI, pp. 605-615, Oct. 1999.

[35] T. Park, T. Tugbawa, D. Boning, S. Hymes, T. Brown, K. Smekalin, and G. Schwartz, "Multi-Level Pattern Effects in Copper CMP," *Proc. CMP Symposium, Electrochemical Society Meeting*, Honolulu, HI, pp. 94-100, Oct. 1999.

[36] T. Tugbawa, T. Park, and D. Boning, "Framework for Modeling of Pattern Dependencies in Multi-Step Cu CMP Processes," *SEMICON West*, July 2000.

[37] L. Liebmann, A. Molless, R. Ferguson, A. Wong, and S. Mansfield, "Understanding Across Chip Line Width Variation: The First Step Toward Optical Proximity Correction," *Proc. SPIE*, Vol. 3051, pp. 124-136, 1997.

[38] P.-Y. Yan and J. Langston, "Mask CD Control Requirement at 0.18 μm Design Rules for 193 nm Lithography," *Proc. SPIE*, Vol. 3051, pp. 164-169, 1997.

[39] T. Park, T. Tugbawa, and D. Boning, MIT/SEMATECH Cu Damascene CMP Characterization Mask, 1998.

[40] B. Stine, D. Boning, J. Chung, D. Ciplickas, and J. Kibarian, "Simulating the Impact of Pattern-Dependent Poly-CD Variation on Circuit Performance," *IEEE Trans. on Semi. Manuf.*, Vol. 11, No. 4, pp. 552-556, Nov. 1998.

[41] Y. Pati, A. Ghanzanfarian, and R. Pease, "Exploiting Structure in Fast Aerial Imaging Computation for Integrated Circuit Patterns," *IEEE Trans. on Semi. Manuf.*, Vol. 10, No. 1, pp. 62-73, Feb. 1997.

[42] V. Axelrad, N. Cobb, M. O'Brien, V. Boksha, T. Do, T. Donnelly, Y. Granik, E. Sahouria, and A. Balasinski, "Efficient Full-Chip Yield Analysis Methodology for OPC-Corrected VLSI Designs," *Proc. International Symposium on Quality Electronic Design*, pp. 461-466, 2000.

[43] S. Zanella, A. Nardi, M. Quarantelli, A Neviani, and C. Guardiani, "Analysis of the Impact of Intra-Die Variance on Clock Skew," *Proc. International Workshop on Statistical Metrology,* pp. 14-17, June 1999.

[44] P. Zarkesh-Ha, T. Mule, and J. Meindl, "Characterization and Modeling of Clock Skew with Process Variations," *Proc. CICC,* pp. 441-444, 1999.

[45] P. Zarkesh-Ha and J. Meindl, "Optimum Chip Clock Distribution Networks," *Proc. International Interconnect Technology Conference,* pp. 18-20, June 1999.

[46] B. Stine and R. Vallishayee, "On the Impact of Dishing in Metal CMP Processes on Circuit Performance," *Proc. International Workshop on Statistical Metrology*, pp. 64-67, June 1998.

[47] B. Stine, V. Mehrotra, D. Boning, J. Chung, and D. Ciplickas, "A Simulation Methodology for Assessing the Impact of Spatial/Pattern Dependent Interconnect Parameter Variation on Circuit Performance," *Proc. IEDM*, Wash. D.C. pp. 133-136, Dec. 1997.

[48] V. Mehrotra, S. Nassif, D. Boning, and J. Chung, "Modeling the Effects of Manufacturing Variation on High-Speed Microprocessor Interconnect Performance," *Proc. IEDM*, San Francisco, CA, pp. 767-770, Dec. 1998.

[49] V. Mehrotra, S. L. Sam, D. Boning, A. Chandrakasan, R. Vallishayee, and S. Nassif, "A Methodology for Modeling the Effects of Systematic Within-Die Interconnect and Device Variation on Circuit Performance," *Proc. DAC*, Los Angeles, CA, pp. 172-175, June 2000.

[50] P. Hofstee, N. Aoki, D. Boerstler, P. Coulman, S. Dhong, B. Flachs, N. Kojima, O. Kwon, K. Lee, D. Meltzer, K. Nowka, J. Park, J. Peter, S. Posluszny, M. Shapiro, J. Silberman, O. Takahashi, and B. Weinberger, "A 1 GHz Single-Issue 64 b PowerPC Processor," *Proc. ISSCC*, pp. 92-93, Feb. 2000.

[51] Virtuoso Layout Editor User's Guide, Cadence, 2000.

[52] Dracula User's Guide, Cadence, 1990.

[53] 3DX Capacitance Extraction Tool User's Guide, IBM, 1998.

[54] S. Srinivasan, *Advanced Perl Programming*, O'Reilly and Associates, Aug. 1997.

[55] Matlab User's Guide, The Mathworks, 1999.

[56] SiCat User's Guide, PDF Solutions, 1999.

[57] Hspice Manual, Avant!, 1999.

[58] Y. Liu, L. Pileggi, and A. Strojwas, "Model Order Reduction on RC(L) Interconnect, Including Variational Analysis," *Proc. DAC,* pp. 201-206, 1999.

[59] D. Ouma, *Modeling of Chemical Mechanical Polishing for Dielectric Planarization*, Ph.D. Thesis, MIT, Nov. 1998.

[60] D. Ouma, D. Boning, J. Chung, G. Shinn, L. Olsen, and J. Clark, "An Integrated Characterization and Modeling Methodology for CMP Dielectric Planarization," *Proc. International Interconnect Technology Conference*, pp. 67-69, June 1998.

[61] J. Silberman, N. Aoki, D. Boerstler, J. Burns, S. Dhong, A. Essbaum, U. Ghoshal, D. Heidel, P. Hofstee, K. Lee, D. Meltzer, H. Ngo, K. Nowka, S. Posluszny, O. Takahashi, I. Vo, and B. Zoric, "A 1.0-GHz Single-Issue 64-bit PowerPC Integer Processor," *IEEE JSSCC*, Vol. 33, No. 11, pp. 1600-1608, Nov. 1998.

[62] B. Lee, D. Boning, D. Hetherington, and D. Stein, "Using Smart Dummy Fill and Selective Reverse Etchback for Pattern Density Equalization," *Proc. Chemical Mechanical Polish for ULSI Multilevel Interconnection Conference (CMP-MIC)*, Santa Clara, CA, pp. 255-258, March 2000.

[63] B. Stine, D. Boning, J. Chung, L. Camilletti, F. Kruppa, E. Equi, W. Loh, S. Prasad, M. Muthukrishnan, D. Towery, M. Berman, and A. Kapoor, "The Physical and Electrical Effects of Metal Fill Patterning Practices for Oxide Chemical Mechanical Polishing Processes," *IEEE Trans. on Electron Devices*, Vol. 45, No. 3, pp. 665-679, March 1998.

[64] S. Nassif, "Design for Variability in DSM Technologies," *Proc. International Symposium on Quality Electronic Design*, pp. 451-454, 2000.

[65] S. Nassif, "Modeling and Forecasting of Manufacturing Variations," *Proc. International Workshop on Statistical Metrology*, pp. 2-10, June 2000.

[66] C. Alpert, A. Devgan, and S. Quay, "Buffer Insertion for Noise and Delay Optimization," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 18, No. 11, pp. 1633-1645, Nov. 1999.

[67] R. Venkatesan, J. Davis, K. Bowman, and J. Meindl, "Optimal Repeater Insertion for N-Tier Multilevel Interconnect Architectures," *Proc. International Interconnect Technology Conference*, pp. 132-134, June 2000.

[68] J. Eble, V. De, D. Wills, and J. Meindl, "Minimum Repeater Count, Size, and Energy Dissipation for Gigascale Integration (GSI) Interconnects," *Proc. International Interconnect Technology Conference*, pp. 56-58, 1998.

[69] V. Mehrotra and D. Boning, "Technology Scaling Impact of Variation on Clock Skew and Interconnect Delay," *Proc. International Interconnect Technology Conference,* June 2001.

[70] J. Meindl, V. De, and B. Agrawal, "Prospects for Gigascale Integration (GSI) Beyond 2003," *Proc. ISSCC*, pp. 124-125, Feb. 1993.

[71] K. Rahmat, O. S. Nakagawa, S.-Y. Oh, and J. Moll, "A Scaling Scheme in Deep-Submicron Processes," *Proc. IEDM*, pp. 245-248, Dec. 1995.

[72] M. Bohr, "Interconnect Scaling -- The Real Limiter to High Performance ULSI," *Proc. IEDM*, pp. 241-244, Dec. 1995.

[73] J. Davis, V. De, and J. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) -- Part I: Derivation and Validation," *IEEE Trans. on Electron Devices*, Vol. 45, No. 3, pp. 580-589, March 1998.

[74] J. Davis, V. De, and J. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) -- Part II: Applications to Clock Frequency, Power Dissipation, and Chip Size Estimation," *IEEE Trans. on Electron Devices*, Vol. 45, No. 3, pp. 590-597, March 1998.

[75] S. Takahashi, M. Edahiro, and Y. Hayashi, "Interconnect Design Strategy: Structures, Repeaters and Materials toward 0.1 μm ULSIs with a Giga-hertz Clock Operation," *Proc. IEDM*, pp. 833-836, Dec. 1998.

[76] K. Yamashita and S. Odanaka, "Interconnect Scaling Scenario Using a Chip Level Interconnect Model," *Proc. Symposium on VLSI Technology*, pp. 53-54, 1997.