Preface

Progress in microelectronics over the last several decades has been intimately linked to our ability to accurately measure, model, and predict the physical properties of solid-state electronic devices. This ability is currently endangered by the manufacturing and fundamental limitations of nanometer scale technology, that result in increasing unpredictability in the physical properties of semiconductor devices. Recent years have seen an explosion of interest in Design for Manufacturability (DFM) and in statistical design techniques. This interest is directly attributed to the difficulties of manufacturing integrated circuits in nanometer scale CMOS technologies with high functional and parametric yield.

The scaling of CMOS technologies brought about the increasing magnitude of variability of key parameters affecting the performance of integrated circuits. The large variation can be attributed to several factors. The first is the rise of multiple systematic sources of parameter variability caused by the interaction between the manufacturing process and the design attributes. For example, optical proximity effects cause polysilicon feature sizes to vary depending on the local layout surroundings, while copper wire thickness strongly depends on the local wire density because of chemical-mechanical polishing. The second is that while technology scaling reduces the nominal values of key process parameters, such as effective channel length, our ability to correspondingly improve manufacturing tolerances, such as mask fabrication errors and mask overlay control, is limited. This results in an increase in the relative amount of variation observed. The third, and most profound, reason for the future increase in parametric variability is that technology is approaching the regime of fundamental randomness in the behavior of silicon structures. For example, the shrinking volume of silicon that forms the channel of the MOS transistor will soon contain a small countable number of dopant atoms. Because the placement of these dopant atoms is random, the final number of atoms that end up in the channel of each transistor is a random variable. Thus, the threshold voltage of the transistor, which is determined by the number

of dopant atoms, will also exhibit significant variation, eventually leading to variation in circuit-level performances, such as delay and power.

This book presents an overview of the methods that need to be mastered in understanding state-of-the-art Design for Manufacturability (DFM) and Statistical Design (SD) methodologies. Broadly, design for manufacturability is a set of techniques that attempt to fix the systematic sources of variability, such as those due to photolithography and CMP. Statistical design, on the other hand, deals with the random sources of variability. Both paradigms must operate within a common framework, and their joint understanding is one of the objectives of this book. The areas of design for manufacturability and statistical design are still being actively developed and an established canon of methods and principles does not yet exist. This book attempts to provide a constructive treatment of the causes of variability, the methods for statistical data characterization, and the techniques for modeling, analysis, and optimization of integrated circuits to improve yield. The objective of such a constructive approach is to formulate a consistent set of methods and principles that allow rigorous statistical design and design for manufacturability from device physics to large-scale circuit optimization.

Writing about relatively new areas like design for manufacturability and statistical design presents its difficulties. The subjects span a wide area between design and manufacturing making it impossible to do justice to the whole area in this one volume. We also limit our discussion to problems directly related to variability, with the realization that the term DFM may be understood to refer to topics that we do not address in this book. Thus, we do not discuss topics related to catastrophic yield modeling due to random defects and particles, and the accompanying issues of critical area, via doubling, wire spreading, and other layout optimization strategies for random yield improvement. These topics have been researched extensively, and there are excellent books on the subject, notably [89] and [204].

Also, with the rapid continuous progress occurring at the time of this writing, it is the authors' sincere hope that many of the issues and problems outlined in this book will shortly be irrelevant *solved problems*. We assume that the reader has had a thorough introduction to integrated circuits design and manufacturing, and that the basics of how one creates an IC from the high level system-oriented view down to the behavior of a single MOSFET are well understood. For a refresher, we would recommend [75] and [5].

The book is organized in four major parts. The first part on *Sources of Variability* contains the three chapters of the book that deal with three major sources of variability: front-end variability impacting devices (Chapter 2), back-end variability impacting metal interconnect (Chapter 3), and environmental variability (Chapter 4). The second part on *Variability Characterization and Analysis* contains two chapters. Chapter 5 discusses the design of test structures for variability characterization. Chapter 6 deals with the statistically sound analysis of the results of measurements that are needed to create rigorous models of variability. The third part is on *Design Techniques*

for Systematic Manufacturability Problems, and deals with techniques of design for manufacturability. Chapter 7 describes the interaction of the design and the lithographic flow, and methods for improving printability. Chapter 8 is devoted to a description of techniques for metal fill required to ensure good planarity of multi-level interconnect structures. The final part on *Statistical Circuit Design* is devoted to statistical design techniques proper: it contains four chapters dealing with the prediction and mitigation of the impact of variability on circuits. Chapter 9 presents strategies for statistical circuit simulation. Chapter 10 discusses the methods for system-level statistical timing analysis using static timing analysis techniques. In Chapter 11, the impact of variability on leakage power consumption is discussed. The final chapter of the book, Chapter 12, is devoted to statistical and robust optimization techniques for improving parametric yield.

This book would not be possible without the generous help and support of a lot of people: our colleagues and graduate students. Several individuals have been kind enough to read through the entire manuscript or its parts, and give the authors essential feedback. Their comments and suggestions have helped us to make this book better. We would like to specifically thank Aseem Agarwal, Shayak Banerjee, Puneet Gupta, Nagib Hakim, Yehea Ismail, Murari Mani, Dejan Markovic, Alessandra Nardi, Ashish Singh, Ashish Srivastava, Brian Stine, Wei-Shen Wang, Bin Zhang, and Vladimir Zolotov. We want to particularly thank Wojciech Maly and Lou Scheffer for reading the entire manuscript and giving us invaluable advice. We thank Denis Gudovskiy for his help with typesetting. Carl Harris, our publisher at Springer, has been a source of encouragement throughout the process of writing. And, of course, this book owes an enormous debt to our families.

Austin, Texas — Michael Orshansky

Austin, Texas — Sani Nassif

Boston, Massachusetts — Duane Boning

July 2007

2

FRONT END VARIABILITY

> There are more things in heaven
> and earth, Horatio, than are
> dreamt of in your philosophy.
>
> William Shakespeare

2.1 INTRODUCTION

One of the most notable features of nanometer scale CMOS technology is the increasing magnitude of variability of the key parameters affecting the performance of integrated circuits [6]. Several taxonomies can be used to describe the different variability mechanisms according to their causes, spatial scales, the particular IC layer they impact, and whether their ability can be described using nonstochastic models. Here we briefly discuss these taxonomies.

The entire semiconductor flow is often partitioned into its front-end and back-end components. The front-end cluster comprises manufacturing steps that are involved in creating devices: implantation, oxidation, polysilicon line definition, etc. On the other hand, the back-end cluster comprises steps involved in defining the wiring of the integrated circuit: deposition, etching, chemical-mechanical polishing, etc. Both front-end and back-end flows exhibit significant variability. In this chapter we concentrate on the front-end variability. It is difficult to say, in a general way, which group of variability contributors dominates. This question can only be answered with respect to a specific concern — overall parametric yield, timing variability. In terms of the resulting timing variability, front-end (device) variability appears to be dominant. For example, for a realistic design, device-caused delay variability contributed close to 90% of the total variability of the canonical path delay [86]. While the exact decomposition of delay variability is design-specific, the device-caused variability is likely to remain the dominant source of path delay

variation, because circuit design practices universally used to reduce the delay of long interconnect lines also help in reducing delay variability due to global interconnect.

It is sometimes useful to distinguish the sources of variability between those related to the issues of manufacturing control and engineering, i.e., extrinsic causes of variation; and those that are due to the fundamental atomic-scale randomness of the devices and materials, i.e., intrinsic causes of variation. The extrinsic manufacturing causes are the more traditional ones and are due to unintentional shifts in processing conditions related to the semiconductor fab's quality of process control. Examples of variability sources in this category include the lot-to-lot and wafer-to-wafer control of oxide thickness growth, primarily determined by the temperature, pressure, and other controllable factors. Historically, scaling made controlling this variability more difficult: while the nominal target values of the key process parameters, such as effective channel length of the CMOS transistors or the interconnect pitch, are being reduced, our ability to improve the manufacturing tolerances, such as mask fabrication and overlay control, is lagging behind [7].

However, the most profound reason for the future increase in parameter variability is that the technology is approaching the regime of fundamental randomness in the behavior of silicon structures. Fundamental intrinsic randomness is due to the limitations imposed by trying to operate devices at the scale at which quantum physics needs to be used to explain device operation and trying to geometrically define materials at the dimensional scale that is comparable to the atomic structure of the materials. In other words, the key dimensions of MOS transistors approach the scale of the silicon lattice distance, at which point the precise atomic configuration becomes critical to macroscopic device properties [8]. At this scale, the traditional descriptions of device physics based on modeling semiconductor with smooth and continuous boundaries and interfaces break down [9].

The primary cases of fundamental device variability are: threshold voltage variation, line-edge roughness, thin film thickness variation, and energy level quantization [10], [11], [12], [13]. For example, because placement of dopant atoms introduced into silicon crystal is random, the final number and location of the atoms that end up in the channel of each transistor is a random variable. As the threshold voltage of the transistor is determined by the number and placement of dopant atoms, it will exhibit a significant variation [14], [15]. This leads to variation in the transistors' circuit-level properties, such as delay and power [16]. Energy quantization will also become a real factor in circuit design. For example, electric noise due to the trapping and de-trapping of electrons in lattice defects may result in large current fluctuations, and those may be different for each device within a circuit. At this scale, a single dopant atom may change device characteristics, leading to large variations from device to device [17]. As the device gate length approaches the correlation length of the oxide-silicon interface, the intrinsic threshold voltage fluctuations induced by local oxide thickness variation will become significant.

For conventional MOSFETs this means that for technologies below 32nm, V_{th} variation due to oxide thickness variation will be comparable to that introduced by random discrete dopants [14], [10]. Finally, line-edge roughness, i.e., the random variation in the gate length along the width of the channel, will become quite noticeable for devices below 50nm, and will be severe at 32nm, also contributing to the overall variability of gate length [12].

The second distinction is based on the spatial scale in which variability of parameters manifests itself. This classification applies to extrinsic variability, as intrinsic variability, by definition, occurs on the scale of a single device. The total variability can be separated into (i) lot-to-lot, (ii) wafer-to-wafer within the lot, (iii) across-wafer, (iv) across-reticle, and (v) within-chip. Different processing steps impact these various spatial scales. The relative magnitudes of each scale depend on the specifics of the process. In general, there tends to be much more between-chip variation across the wafer compared to wafer-to-wafer variation within the lot [49].

For the circuit designer's sake, the primary distinction is between chip-to-chip (interchip) and within-chip (intrachip) variability. Historically, within the chip the variation of the parameters could be safely neglected in digital circuit design (analog designers have been concerned with matching for a long time). The patterns of variability are changing, however. For 130nm CMOS technologies, the percentage of the total variation of the effective MOS channel length that can be attributed to intrachip variation can be up to 35% [18]. A useful distinction that relates to within-chip variability is based on similar structure variability and dissimilar-structure variability [49]. Variability between similar structures arises due to the across-wafer and across-reticle variability that every chip experiences. Variability between dissimilar structures may be due to (i) the differences in processing steps, for example, different masks are used in dual threshold voltage processes for making devices with low and high Vth, and (ii) different dependencies of process conditions to variations in layout orientation and density, for example, orientational dependence in lithography or micro-loading in resist and etch.

The increase of intra-chip parameter variation is caused by the emergence of a number of variation-generating mechanisms located on the interface between the design and process. For example, one of the major contributors to the variation of the effective channel length is the optical proximity effect. As the transistor feature size gets smaller compared to the wavelength of light used to expose the circuit, the light is affected by diffraction from the features located nearby. As a result, the length of the final polysilicon line becomes dependent on the local layout surroundings of each individual transistor.

Another source of large intrachip parameter variation is the aberrations in the optical system of the stepper used to expose the mask. These aberrations lead to predictable systematic spatial variation of the MOS gate length across the chip. For interconnect, an important source of variability is the dependence of the rate of chemical-mechanical polishing (CMP) on the underlying density of interconnect. The most significant problems that may arise when polishing

are dishing and erosion, which happen when some areas of the chip are polished faster than others. In dishing, the metal (usually copper) is "dished" out of the lines. Erosion happens when some sections of the interlevel dielectric are polished faster than others.

An important distinction that is often misused is between the stochastic (random, statistical) variability and the systematic (deterministic) variability mechanisms. The confusion stems from not distinguishing the actual mechanism by which variation is generated from one's ability to predict the value of a variable deterministically (and thus analyze, correct, and compensate for it). A nonstatistical (deterministic) description does not make a reference to the variance of a process parameter, but only to its mean value. For example, a well-specified nonuniform temperature profile affects the entire wafer and is thus systematic to the process engineer who can measure it and observe that the same profile affects each wafer in an identical way. Let us suppose that the process engineers cannot correct this temperature nonuniformity. To a circuit designer, this source of variability will appear statistical: the placement of each die on the wafer is unknown and cannot be utilized. There is no way by which the circuit designer can deterministically describe the values of temperature affecting each die, and thus only a statistical description is possible. (The statistical variable used can be spatially correlated, however.) In summary, the importance of the distinction is that we must treat random and systematic variations differently. While the systematic variations are modeled and accounted for in the design flow, the random variations are handled either through worst-case margining or parametric yield optimization methods.

It is interesting to inquire about the future trends that the variability components will exhibit. Is variability going to grow dramatically or remain under control? In general it is quite difficult to predict the magnitude of variability that will be characteristic of future processes, or even make reliable generalizations across the current processes. However, several trends appear quite certain. The threshold voltage variability will rise driven by the increased contribution of the random dopant fluctuations. At the limit of scaling, below 22nm nodes, the oxide thickness variation and line edge roughness are likely to be substantial contributors to the variability budget. Until new lithography solutions are adopted in place of the current 192nm exposure systems, gate length (L_{gate}) variability due to lithography is bound to remain problematic. For other variability mechanisms, the future is less predictable, as ways to improve control are continuously developed. Figure 2.1 shows a large increase in 3σ variation of effective transistor length (L_{eff}), oxide thickness (T_{ox}), threshold voltage (V_{th}), interconnect width (W) and height (H), and dielectric constant (ρ) [73]. These predictions should be interpreted cautiously, since the ability to control specific sources can change in the future, for better or for worse.

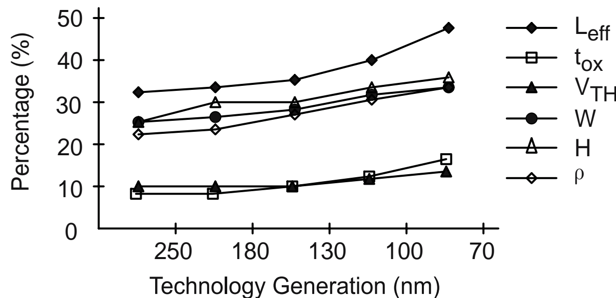


Figure 2.1 The 3σ parameter variation increases as a result of scaling (Reprinted from [73], ©2000 IEEE)

2.2 VARIABILITY OF GATE LENGTH

2.2.1 Gate Length Variability: Overview

Variability in the gate length of the MOS transistor is extraordinarily important for multiple aspects of IC performance and design. This parameter is known as "critical dimension" in the manufacturing community because it defines the minimum feature size of a technology. Electrically, gate length and a related parameter, known as effective channel length (L_{eff}) strongly impact the current drive, and therefore the speed, of the circuit. There are several ways to define the effective channel length; here we take it to be equal to the gate length minus the under-diffusions of the source and drain regions. In the discussion that follows we will adopt the term L_{gate} uniformly. Another term that sometimes appears is the critical dimension (CD) that lithographers use to refer to L_{gate}.

Transistor leakage current is an exponential function of L_{gate}. Because of this exponential dependence, variation of L_{gate} is greatly amplified in its impact on leakage. The growth of power consumption has led to a situation in which many chips are power-limited. As a result, L_{gate} variability leads to a large parametric yield loss. Because this loss occurs primarily in fast frequency bins which are the most profit-generating bins, L_{gate} variability is economically very costly. It has been estimated by one major semiconductor company that a reduction of 1nm of the standard deviation (σ) of L_{gate} would result in an additional earning of \$7.5/chip for a high-end product [19]. For future technologies, this cost of variability in L_{gate} is likely to be much higher. The ITRS Roadmap requires total L_{gate} variation (3σ) to remain under 10%; however, for technologies beyond 45nm node, a manufacturable solution is still unknown [19].

A large number of processing steps and modules have an impact on effective channel length. Those include the mask, the exposure system, etching, the spacer definition, and implantation of source and drain regions. Factors that

Table 2.1 Summary of contributions to L_{gate} variability from different processing modules [21]

Process Step	**Source of Variability**
Wafer	Flatness, reflectivity, topography
Reticle	CD error, defects, edge roughness, proximity effects
Stepper	Aberrations, lens heating, focus, leveling, dose
Etch	Power, pressure, flow rate
Resist	Refractive index, thickness, uniformity, contrast
PEB	Temperature, uniformity, time, delay
Environment	Amines, humidity, pressure
Develop	Time, temperature, dispense, rinse

contribute to the variability of the polysilicon gate width are the dominant contributors to L_{eff} variability [20]. There are also multiple causes in the manufacturing sequence that contribute to overall L_{gate} variation. Table 2.1 provides a fairly exhaustive list of such causes, most of them primarily interesting to process engineers. While the complete list of causes in Table 2.1 is quite extensive, error decomposition indicates that the primary ones include reticle mask errors, variations in scanner/stepper illumination, lens aberrations, post-etch bake (PEB) temperature non-uniformity, and plasma etch rate non-uniformity [22].

From the designer's point of view most of these variability patterns are random. However, at the process level, continuous improvement of statistical metrology and the use of techniques for uncovering complex statistical dependencies have shown that much of the variability in the lithographic part of the sequence is systematic. Other variations acting across the wafer due to the lack of uniformity in temperature, or the non-uniformity of film thickness, may also be highly systematic, at the process level [22].

Similar to other components of variation, linewidth variation can be decomposed into chip-to-chip and within-chip components. The within-chip component is often termed *across-chip linewidth variation (ACLV)*. The chip-to-chip component can be further decomposed into contributions from the lot-to-lot, wafer-to-wafer, and within-wafer components. Slow-changing, long-term fluctuations of the process may lead to lot-to-lot variations. Variations in etch or resist bake may introduce wafer-to-wafer variations. Within-wafer effects may be due to the radial variations in the photoresist coating thickness or etching.

ACLV is primarily determined by systematic effects due to photolithography and etching. Again, a multitude of factors may contribute, including: stepper induced variations (illumination, imaging nonuniformity due to lens aberrations), reticle imperfections, resist induced variations (coat non-uniformity, resist thickness variation), and others. L_{gate} variability within a reticle field exhibits a strong systematic spatial dependence which is primarily due to lens aberrations [18]. The scaling of lithographic features makes the lens

aberrations even more severe by forcing the operation of the illumination system at the optical resolution limit. The variability patterns due to aberrations are highly predictable at the level of the reticle field, and can be accurately described by distinct 2D surfaces. Finally, there also exists an interaction between the global lens aberration and the local layout pattern-dependent nonuniformities due to proximity, which contributes to the overall variability budget. We now discuss two major contributors: photolithography and etch.

2.2.2 Contributions of Photolithography

The delayed introduction of new lithographic processes based on the 157nm wavelength of light, has forced the last several technology generations to use the older technology based on 193nm light. To continue scaling the features, imaging systems had to rely on lower values of k_1, the parameter that is a metric of lithography aggressiveness. The k_1 coefficient is defined as $k_1 = \frac{NA}{\lambda}CD$, where CD is the critical dimension of the feature being printed, λ is the wavelength of light, and NA is the numerical aperture of the lens. Over the years, the value of k_1 has decreased from about 1 to nearly 0.5. With low k_1 imaging, *image distortion* during photolithography is a major contributor to across-chip linewidth variation. It also leads to other shape distortions. The effect of low k_1 is that the optical system has a low-pass filter characteristic, filtering out the high-frequency components of the reticle features. This behavior results in several major types of distortions: linewidth variation (proximity effect), corner rounding, and line-end shortening [22]. These are all systematic behaviors highly dependent on design layout characteristics. The essentials of photolithography relevant to printability are further discussed in Chapter 5.

Proximity effect refers to the dependence of the printed CD on its surrounding. In this simplest 1-D case, the main dependence is on the distance to the nearest neighbor, or equivalently, the pitch. Depending on the proximity of the neighbors, polysilicon features can be classified as isolated or dense (nested). The dependence of linewidth on pitch is determined by the type of the photoresist used. A typical dependence of printed linewidth on the pitch is shown in Figure 2.2.

Line shortening refers to the reduction in the length of a rectangular feature. This effect is due to factors that include diffraction, the rounding of the mask patterns themselves, and photoresist diffusion. At low k_1 imaging, diffraction is a major reason, and with smaller CD, line shortening grows rapidly. Corner rounding is another type of image distortion, which occurs because the high-frequency components of the corner are filtered, producing a smoothed-out pattern. This has a large impact on the gate width of the transistor if the polysilicon gate is laid out very near the L-shaped active region of the transistor. Because of the rounding of the corners of the L-shaped region, the effective gate width depends on the relative position of the gate and active regions. Line shortening and corner rounding are illustrated in Figures 2.3 and 2.4 respectively.

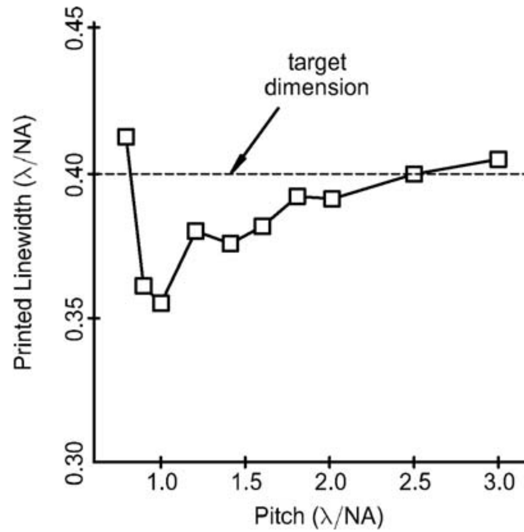
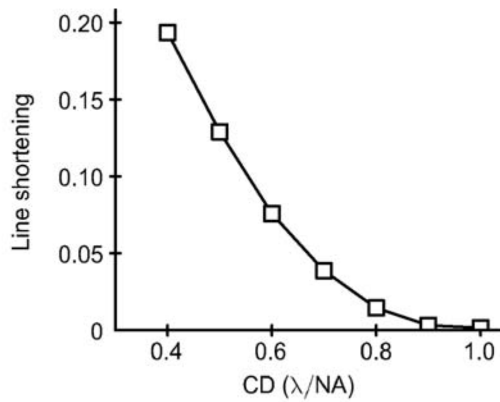


Figure 2.2 A typical dependence of linewidth on the proximity to the neighboring polysilicon lines (Reprinted from [136], ©2001 SPIE)



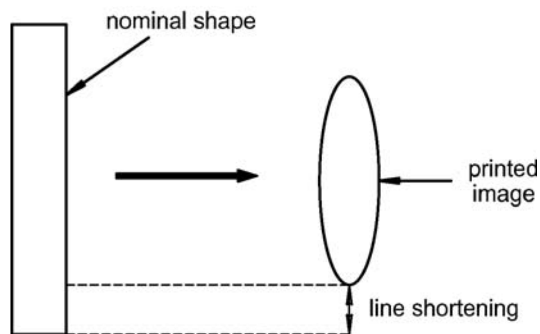


Figure 2.3 Line shortening (Reprinted from [136], ©2001 SPIE)

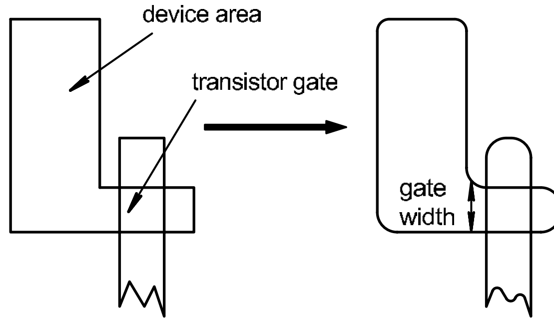


Figure 2.4 Corner rounding (Reprinted from [136], ©2001 SPIE)

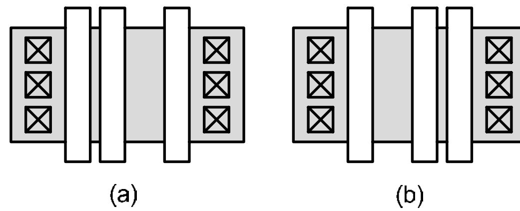
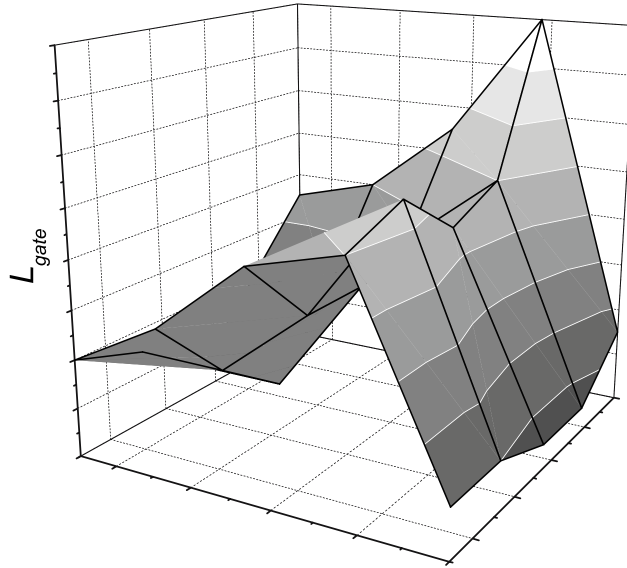


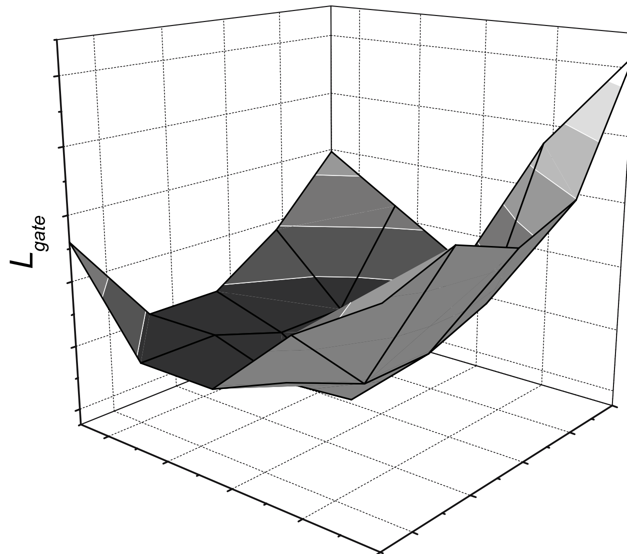
Figure 2.5 Data shows that linewidth depends on the relative positions of the neighbors and exhibits asymmetry. Linewidth in layout pattern (a) is predictably different from that in pattern (b)

Lens aberrations may lead to significant systematic spatial non uniformity of L_{gate} over the reticle field. The spatial variation across the reticle can be as high as 12%, for a technology with $L_{gate} = 130$nm. Depending on the placement of a circuit, such as a ring oscillator, within the die its speed could vary by almost 15% [18]. The spatial L_{gate} maps that characterize the variations also depend on the local neighborhoods of the polysilicon features: dense and isolated features will exhibit different spatial profiles, indicating statistical interaction between the global lens aberrations and the pattern-dependent optical proximity effect. Lens imperfections also lead to predictable L_{gate} bias between the gates that are oriented vertically or horizontally in the layout. Finally, the coma effect leads to an anisotropy of multifingered layouts: the relative position of the surrounding gates, i.e., the neighbor being on the left vs. right, exerts a predictable impact on the final L_{gate}, as in Figure 2.5. This anisotropy also leads to spatial across-reticle maps L_{gate} that are distinctly different, as in Figure 2.6. These differences are systematic, i.e., predictable, which is supported by rigorous analysis of variance.

Another factor that has to be taken into account is the increased mask error factor (MEF), also known as mask error enhancement factor (MEEF). In projection photolithography, features on photomasks are scaled exactly onto the wafer by the demagnification of the projection optics ($1/M$). At large k_1, the mask errors arising due to the inability to ideally place the features



(a)



(b)

Figure 2.6 The systematic spatial L_{gate} variation across the reticle field (a) The spatial profile for a gate with the nearest neighbor on the left and a moderately spaced neighbor on the right (b) The spatial profile for a gate with the nearest neighbor on the right and a moderately spaced neighbor on the left

on the mask are scaled by the same demagnification factor. For example, if M=5, then the 20nm error in the mask feature placement will result in only a 4nm printed CD error. However, at low k_1 imaging, for $0.5 < k_1 < 0.8$, the beneficial effect of demagnification on the mask error is reduced. Effectively, the mask error gets magnified and the degree of such error magnification is described by the mask error factor:

$$\Delta CD_{resist} = MEF * \frac{\Delta CD_{mask}}{M} \tag{2.1}$$

While this particular contributor to L_{gate} variability has always been present, it has recently taken on increased importance. The primary cause of MEF is degradation of image integrity, e.g., the loss of image shape control (due to such factors as lens aberrations, defocus, exposure, partial coherence), and photoresist processing at low k_1 [144]. Measurements show that for a given process, MEF increases rapidly for small feature printing, as in Figure 2.7. MEF is a strong function of defocus and exposure errors. Defocus is the vertical displacement of the image plane during illumination. Exposure errors are due to differences in energy delivered by the illumination system, and other process errors that behave similarly to exposure errors. MEF also depends on local layout density: it is higher for nested lines and spaces than for sparse lines [23]. The result of the increased value of MEF is that the mask placement errors contribute a growing amount to the overall L_{gate} variability. However, the dependence of MEF on the design attributes can be used to increase the process window and reduce the impact of mask errors on L_{gate} variability.

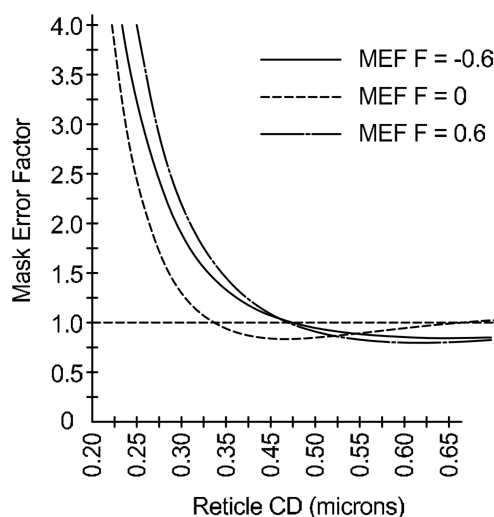


Figure 2.7 Mask error factor grows rapidly at smaller linewidths. It also depends strongly on defocus (Reprinted from [23], ©2003 SPIE)

2.2.3 Impact of Etch

The impact of etching nonuniformity on the overall linewidth budget can be comparable to the contribution of photolithography [20]. Etching nonuniformity manifests itself as variability of etching bias, which is the difference between the photoresist and etched polysilicon critical dimensions. From the designer's perspective, the variation of etching bias as a function of layout pattern density is the most important component. This dependence can be classified into three groups: micro- and macroloading, and aspect-ratio-dependent etching. In aspect-ratio-dependent etching, the variation of linewidth is dependent on the distance to nearby features [26]. The biases due to photolithography and etching processes are additive.

Microloading and macroloading are driven by a common physical mechanism. The variation in the layout features can increase or decrease the density of the reactant. In microloading, the etching bias for the same drawn features will depend on the local environment, with the range of influence of different patterns being 1–10 mm. Significant microloading can occur in places where there are abrupt changes in density, e.g., near scribe-lines, test and in-line diagnostic chips, and near the wafer edge [25]. In macroloading, the etching bias is determined by the average loading across the wafer [25]. Macroloading is a problem for technologists and process engineers, particularly for fabs that manufacture different types of ICs, e.g., logic, DRAM, and gate arrays.

2.2.4 Line Edge Roughness

Despite the limitations of the patterning process discussed so far, the existing photolithographic processes are capable of producing a consistent poly line edge. As the devices are scaled below 50nm, the random variation in the gate length along the width of the gate will become quite noticeable making gate length variation control even more difficult, and its impact will become severe below 32nm [12]. Line edge roughness is the local variation of the edge of the polysilicon gate along its width. The reasons for the increased LER in the future processes include the random variation in the incoming photon count during exposure and the contrast of the aerial image, as well as the absorption rate, chemical reactivity, and the molecular composition of resist [27], [49]. Figure 2.8 shows the randomness of the line edge through several steps of the via hole fabrication process.

Line edge roughness has an impact on all the main electrical device characteristics: the drive current, off-current, and the threshold voltage. The easiest way to characterize the line edge roughness is to compute its variance. For example, in a 193nm process, the total variation due to LER has the standard deviation of $3\sigma_{LER} = 6.6 - 9$nm, measured on a polysilicon line with $L_{gate} = 110$nm. However, the knowledge of the variance of LER is insufficient to properly predict at least some parameters, for example, the leakage current. The current value also depends on the spatial frequency profile of the

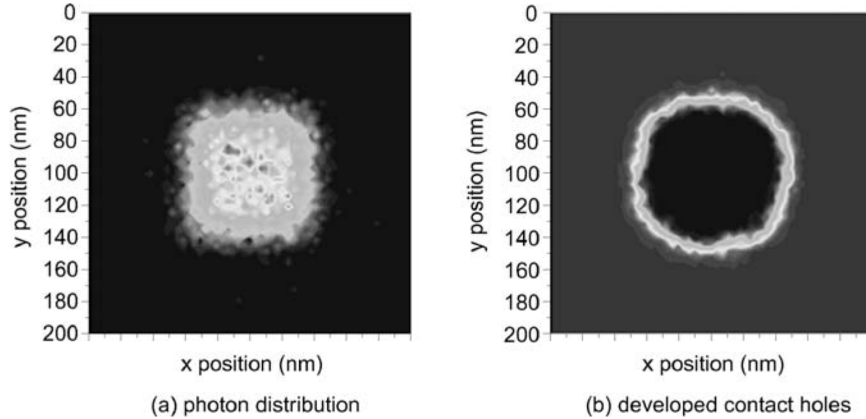


Figure 2.8 Simulation of the exposure and development of a via hole with extreme ultraviolet lithography (Reprinted from [51], ©2003 SPIE)

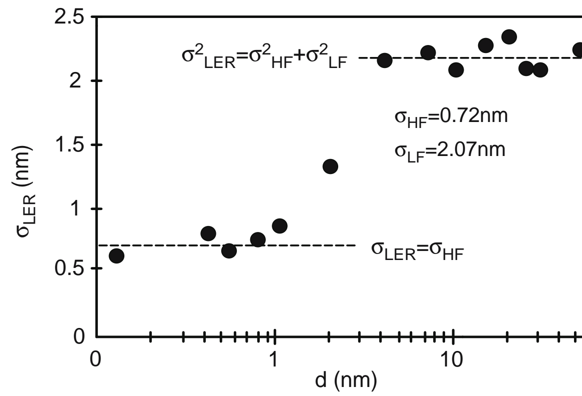


Figure 2.9 Variance of line edge roughness depends on gate width. The variance increase saturates beyond about 0.3μm (Reprinted from [12], ©IEEE 2002)

local roughness. LER measurements show that the edge profile exhibits both smooth, slow changing (low-frequency) and high-frequency types of variation [12]. For this reason, measurements show that there is strong dependence of edge variance on the polysilicon gate width. Once the gate width is greater than ~0.3μm, the variance does not increase any more, as in Figure 2.9. Thus, capturing only the variance ignores the spatial frequency profile of LER and fails to predict the variance dependence on the length of the measured line. A complete description would include the characterization of the spatial frequency of the LER [28].

A model that can be more physically helpful relies on extracting only two additional parameters, the correlation length (ξ) and the roughness exponent

(α). Correlation length is a measure of the length after which the segments of the polysilicon edge can be considered uncorrelated. The roughness exponent is a measure of the relative contribution of the high frequency component to LER. Higher values of α correspond to smoother lines with less high-frequency variation. When $\alpha \sim 1$, the profile exhibits a periodic behavior. Experiments show that for the 193nm process, the correlation length is about 33nm, and is relatively insensitive to aerial image quality. The roughness exponent increases slightly with decreasing aerial image contrast, suggesting that at high contrast imaging the contribution of the high frequency is greater. The variance value saturates beyond about 10ξ, i.e., at about 0.3µm. To assess the device-level impact of line edge roughness, we need to translate it into device width roughness, which results from local variation of both polysilicon edges. Experimental measurements show that the roughness of the two edges can be considered uncorrelated. Then, for gate width $W > 0.3$µm, L_{gate} variance caused by line edge roughness can be approximated as $2\sigma^2_{LER}$.

Most of the experimental evidence suggests that the $3\sigma_{LER}$ is in the range of 5–6nm. These numbers are quite consistent among many companies and across several technology generations. The reported values of the correlation length, however, range much more widely: 10–50nm [48]. 2D device simulations indicate that below 32nm, LER will have a significant impact on V_{th} uncertainty and will lead to a large increase in leakage current. Assuming $3\sigma_{LER} = 6$nm and the correlation length of 20Å, simulations show that in a device with $L_{gate} = 30$nm, the variation in threshold voltage, at low V_{ds}, is $\sigma_{V_{th}} = 8$mV. In a device with $L_{gate} = 50$nm, the variation in threshold voltage is less severe: $\sigma_{V_{th}} = 2.5-5$mV [29][48]. It is instructive to compare the impact of LER on threshold voltage variability with that of random dopant fluctuation, discussed later. At $L_{gate} = 30$nm, random dopant fluctuation will lead to $\sigma_{V_{th}} = 38$mV, approximately, making the impact of LER on threshold voltage uncertainty comparatively small. For nonminimum width devices, the variation in threshold voltage is smaller: $\sigma_{V_{th}}$ varies as $1/(W_{eff})^{0.5}$. Figure 2.10 investigates the dependence of $\sigma_{V_{th}}$ and of leakage increase on L_{eff} [29]. It is clear that below 45nm, line edge roughness does lead to a significantly increased mean leakage current.

2.2.5 Models of L_{gate} Spatial Correlation

For the purpose of modeling of intrachip variation of L_{gate}, a model based on spatial correlation is used. Indeed, it is reasonable to believe that two transistors nearby will be affected by any source of variation in a similar way, leading to correlation. Moreover, this correlation should decrease with the increasing distance between the two transistors. This is the foundation behind the standard Pelgrom model [30]. The form of the correlation function and the value of the correlation length are determined empirically. One possible correlation function is of the form [31]:

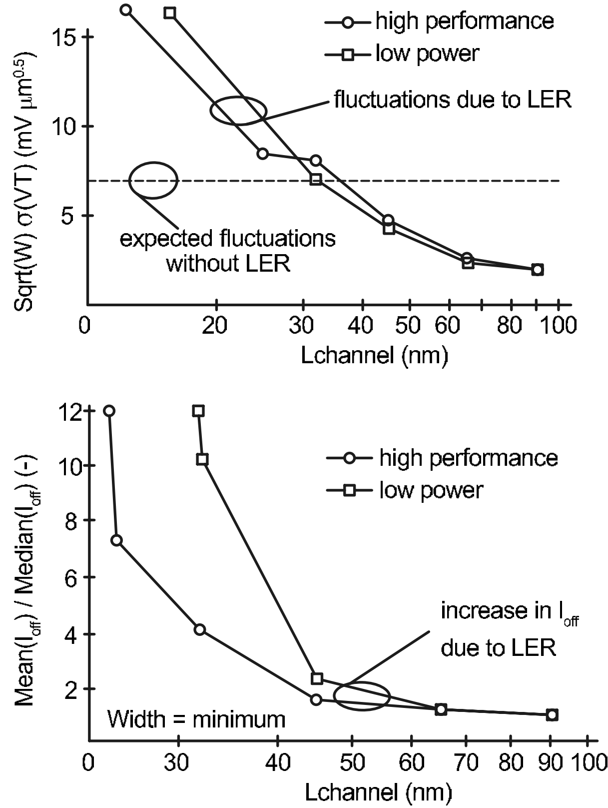


Figure 2.10 LER will have a growing impact on V_{th} uncertainty and electrical characteristics (Reprinted from [29], ©IEEE 2002)

$$Var(\Delta CD_d) = 2\,Var(CD)\left(1 - \exp\left(\frac{-d}{dl}\right)\right) \tag{2.2}$$

where *Var(CD)* is the total CD variance of a single device, and *dl* is a characteristic distance for a particular technology.

Discussions of spatial correlation are often confounded with the issue of systematic spatial variation. The term "systematic" variation has a fair amount of ambiguity. From the point of view of statistics, "systematic" variation refers to phenomena characterized by the difference in mean values of certain measures. Systematic variation is synonymous with "deterministic" and can be described by functional forms. This is in contrast to random, or stochastic, variability. From an engineering point of view, naming a certain variability pattern "systematic" seems to be justified only if corrective actions can be taken. What may be "systematic" variability from the point of view of process engineers, may not be so from the point of view of circuit designers. For example, across-wafer and across-field CD variations exhibit spatial trends

that appear systematic to the process engineer. Thus, process control can be used to characterize, compensate, and thus eliminate these systematic dependencies. If the data is analyzed now, after the removal of the above components of systematic variation, one finds that the magnitude of spatial correlation that was apparently present in the data is significantly reduced [22].

What if such ideal process control is not implemented? A circuit designer has no way of modeling this variability, except in a statistical sense. Systematic is equivalent to functionally modelable. But to a circuit designer facing a population of chips with different CDs, the above variability appears stochastic. While it is stochastic, it is, at the same time, correlated. The description utilizing spatial correlation is useful even though the spatial correlation is in reality due to a systematic nonstationary structure of the data. It can be noted that the famous Pelgrom model is also based on similar reasoning. In this model, the long-range radial wafer-level variation is clearly systematic. But because of the unknown placement of a die on the wafer, it manifests itself to designers as an additive stochastic component with a long correlation distance [30].

2.3 GATE WIDTH VARIABILITY

For nonminimum size transistors typically used in logic gates, variability in transistor width has a negligible impact on performance parameters. However, for minimum width transistors, width variability is substantial. Mask alignment is a traditional source of width variability. Still, it is primarily because of two reasons. The first is grounded in photolithography. The gate is defined by the overlap between the polysilicon and diffusion layers. Many standard cells are laid out in such a way that the polysilicon gate makes a corner in a close vicinity of the diffusion layer, as in Figure 2.11. As we learned in the previous section, subwavelength lithography introduces image distortions and exhibits features of a low-pass filter when printing features with sharp corners. In this case, corner rounding leads to the reduction of the effective width of the transistor. A very similar situation takes place due to diffusion layer rounding.

Another source of gate width variability is due to the planarization steps involved in producing device isolation based on shallow trench isolation (STI) technology. STI is the dominant isolation technique for deep submicron technologies, favored for its excellent latch-up immunity, low junction capacitance, and sharp vertical edges [32]. STI is performed with a damascene process similar to the one used in copper metallization processes. First, a protective layer of oxide and a layer of thicker nitride on the surface of silicon are deposited. An isolation mask is used to define the trenches. The nitride is patterned and anisotropically etched into the silicon substrate, producing a trench with sharp vertical walls. A reactive ion etch (RIE) is used to etch the silicon trenches. The trench is then filled with oxide, producing an isolation between

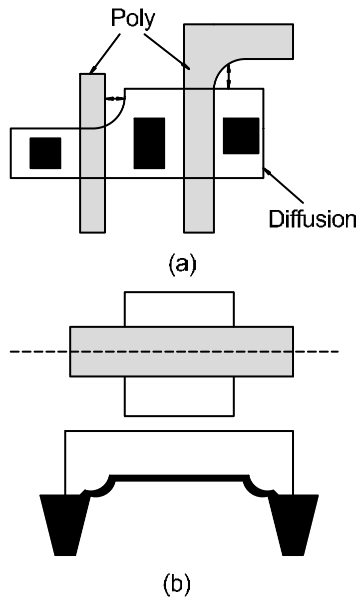


Figure 2.11 The two contributors to gate width variability: (a) corner rounding on the poly and active layers, and (b) the impact of CMP used in shallow trench isolation Courtesy of N. Hakim [31]

the neighboring devices. Now, however, the oxide has to be removed, so that the oxide forming the STI and the silicon of the active areas are coplanar [33]. The planarization is performed using chemical-mechanical polishing (CMP) that removes the material using a combination of mechanical pressure and chemical action. As in the case of metal planarization, the rate of removal depends on the material and on the underlying pattern density, i.e., the layout. The wide trenches experience dishing, and thus are lower than the active area silicon. The planarity can be improved by using dummy fill features as well as imposing new design rules on active area density [33]. However, because of the limitations of these control schemes, there is residual nonuniformity in the alignment of silicon and oxide areas. Typically, extra silicon is removed near the STI-device interface. This effectively reduces the width of the transistor due to a nonvertical boundary. While for large widths this effect can be ignored, it is nonnegligible for small devices.

2.4 THRESHOLD VOLTAGE VARIABILITY

The threshold voltage of a MOS transistor is determined by several device characteristics, including the material implementing the gate (typically, highly doped polysilicon), the thickness of the dielectric film (typically, silicon dioxide), and the concentration and the density profile of the dopant atoms in

the channel of the transistor. As a result, the variations in oxide thickness, implantation energy and dose, and the substrate doping profiles will lead to the variation in threshold voltage (V_{th}). Historically, all these effects jointly resulted in 3σ V_{th} variation of less than 10% of the nominal value [73]. Also, because all the above variation sources exhibited variability primarily on the chip-to-chip scale, intrachip V_{th} variation was inconsequential, at least for digital designs. (Analog designers have always been concerned with the problem of matching the threshold voltages of transistors in amplifiers, comparators, and other circuits that require good matching). With the continuing scaling of MOS dimensions, a radically different problem of V_{th} variation due to random dopant fluctuation (RDF) has emerged. Figure 2.12 shows the example of the distribution of threshold voltage from a 65nm CMOS process.

Placement of dopant atoms into the channel is achieved via ion implantation. Implantation and the subsequent activation through annealing are such that the number and placement of atoms implanted in the channel of each transistor is effectively random. Because the threshold voltage of the transistor is determined by the number and location of dopant atoms, it also exhibits a significant variation. Figure 2.13 shows the randomized placement of dopant atoms in the channel of the 50nm MOSFET. The phenomenon of random dopant fluctuation truly belongs to the class of fundamental atomic scale randomness, with precise atomic configuration being critical to macro-

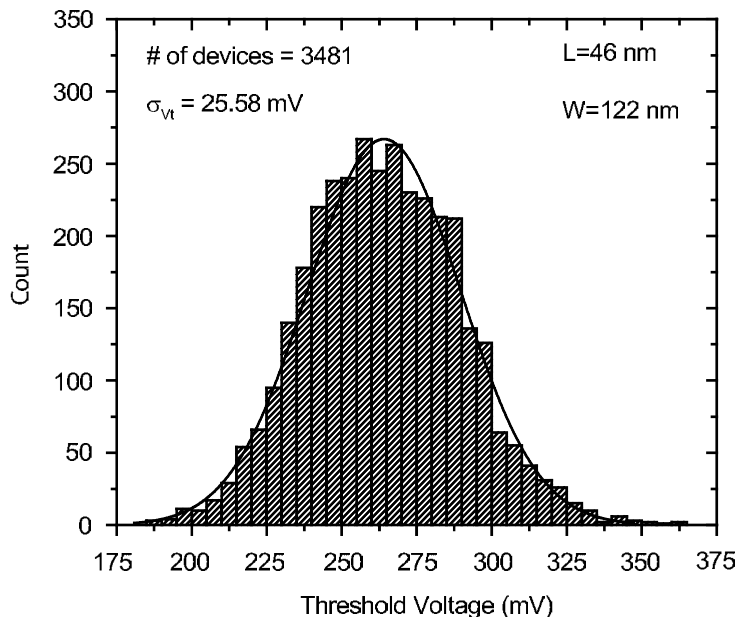


Figure 2.12 Distribution of the n-channel FET threshold voltage from a 65nm CMOS process

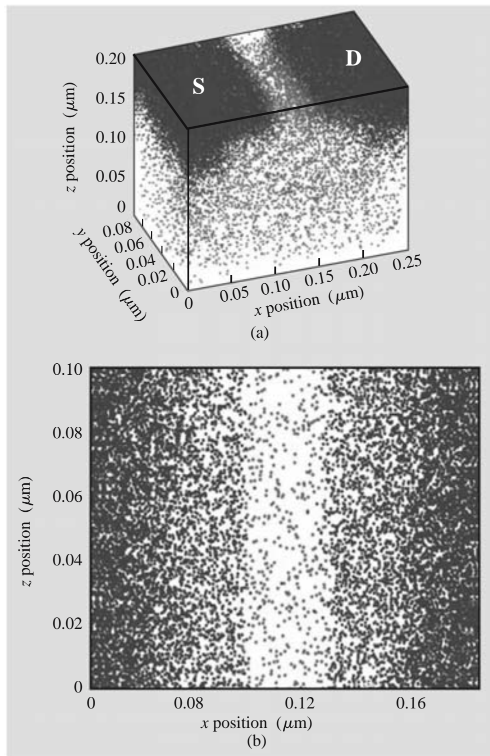


Figure 2.13 The random placement of dopants also impacts the definition of the source and drain regions, leading to the variation of source and drain capacitance and resistance (Reprinted from [49], ©2006 IBM)

scopic properties of devices [34]. The description that models semiconductors with smooth, continuous boundaries and interfaces breaks down [9], and has to be supplemented. Because of this discreteness and the stochastic nature of the implantation process, the location of the dopant atoms will vary from transistor to transistor. At the same time, because the number of the dopant atoms is getting smaller, the variation of the number of dopants around a certain mean value becomes greater.

A theoretical model that predicts the amount of threshold voltage uncertainty can be constructed via a 3D analysis of the distribution of impurities in the silicon substrate [35]. The considered region is equal to a parallelepiped with the depth equal to the average depth of the depletion layer, X. A model divides the entire area into a number of cubes with the edge of length X. Given the average number of impurities, M, in a cube of size X^3, the actual number of impurities, m, is described as following the Poisson distribution:

$$P(m) = m^M e^{-M}/m! \qquad (2.3)$$

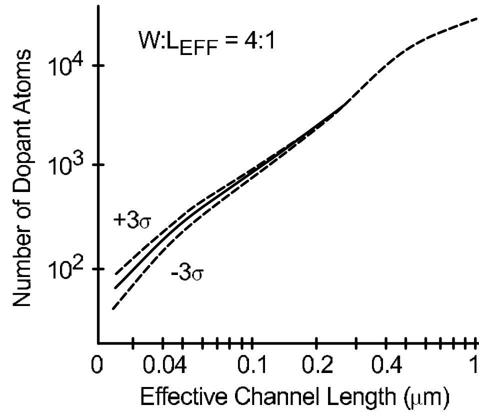


Figure 2.14 The number of dopant atoms in the channel is getting smaller, increasing the relative uncertainty in the actual number (Reprinted from [36], ©2001 IEEE)

Given the properties of the Poisson distribution, if the mean number of dopants is M, the standard deviation of the number of dopants is $M^{1/2}$. It has been empirically found that the mean number of dopant atoms in standard bulk CMOS devices has been decreasing roughly in proportion to $L_{eff}^{1.5}$. Since the mean number of dopant atoms that are placed in the channel at the end of the implantation and activation processes rapidly decreases, the normalized uncertainty (σ/μ) in the number of atoms grows as $1/\sqrt{M}$. Figure 2.14 shows the variance of the number of dopant atoms for different values of the effective channel length.

We now need to model the impact of dopant number uncertainty on the threshold voltage itself. Analytical models exist, and are typically based on the percolation models for establishing a path from source to drain [17], [35], [36]. Such analytical models are indispensable in providing an intuition for the general dependence of the uncertainty on device parameters.

The precise impact of this uncertainty in the number and the placement of dopant atoms on the threshold voltage depends heavily on the specifics of the doping profile used in a MOSFET. Because of the dependence on doping profile, numerical simulations often must be used. For each lattice site, the program computes a probability of it being a dopant, which can be found from the continuum doping concentration. This can be done for the entire substrate and for any doping profile. Then, at each site a dopant atom is randomly placed in accordance with the computed probability [36], and the device electrical properties are analyzed. These numerical simulations allow a look at the magnitude of V_{th} uncertainty for MOSFETs at the limits of scaling. For a device with the 25nm gate length it is predicted that $\sigma_{V_{th}} = 7 \sim 10/\sqrt{W}\text{mV}\cdot\mu\text{m}^{1/2}$. Even if a retrograde doping profile is selected, which is optimal from the point of view of V_{th} uncertainty, the magnitude of uncertainty would remain at $\sigma_{V_{th}} = 5/\sqrt{W}$

mV·μm$^{1/2}$ [36]. More accurate models that take into account the quantum confinement indicate that the uncertainties can be about 24% higher than stated above [34].

Based on these numerical simulations, an empirical model has been developed [14]. It is convenient to designers because it compactly captures the dependence of the V_{th} sigma on several device parameters:

$$\sigma_{V_{th}} = 3.19 \times 10^{-8} \frac{T_{ox} N_A^{0.4}}{\sqrt{L_{eff} W_{eff}}} \tag{2.4}$$

where T_{ox} is the oxide thickness, N_A is the doping density, L_{eff} and W_{eff} are the effective channel length and width. From the design perspective, the important factor in this model is the inverse dependence of the standard deviation of V_{th} on the square root of the transistor width, and thus area. Because of this dependence, the uncertainty (measured in the standard deviation of V_{th}) of large-width devices will be much smaller than that of minimal-width devices. Figure 2.15 presents measurements of $\sigma_{V_{th}}$ for different values of gate area. It can be seen that the data is consistent with the behavior predicted by the model.

All in all, the wide devices used in high-performance logic may have a few extra millivolts of variation, an insignificant amount. The problem is

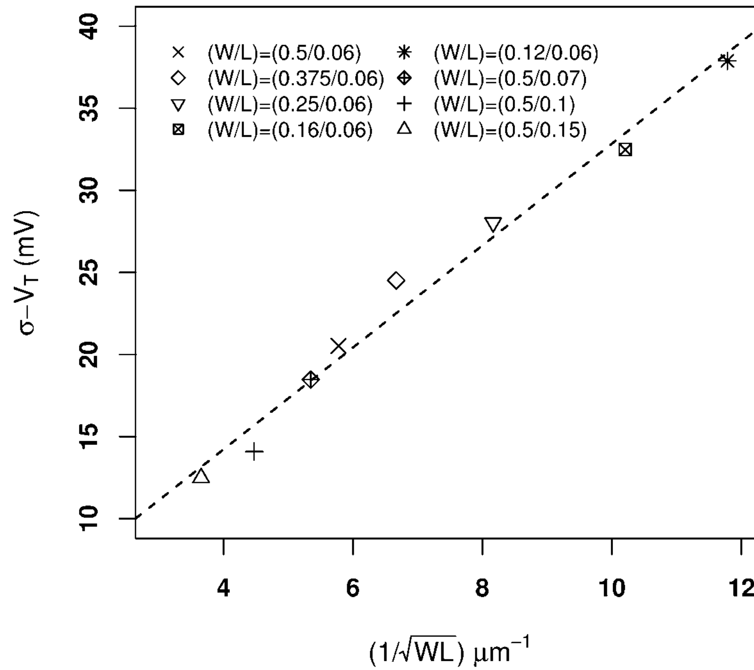


Figure 2.15 Measurements of $\sigma_{V_{th}}$ for different gate geometries

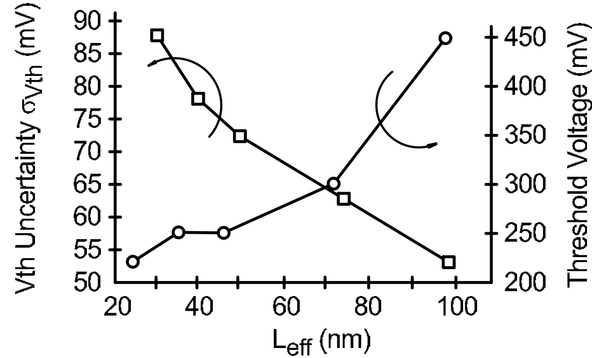


Figure 2.16 While the nominal V_{th} gets smaller, the uncertainty in V_{th} increases

absolutely severe for SRAM designs that rely on minimum-width transistors, which may have $\sigma_{V_{th}} = 40$mV. The magnitude of V_{th} uncertainty due to RDF makes it one of the most difficult problems facing CMOS scaling, and especially, SRAM scaling. A more detailed analysis of the impact of RDF on SRAM is presented later in the book. Figure 2.16 shows the projected magnitudes of $3\sigma_{V_{th}}$ together with the nominal saturated threshold voltage for several values of L_{gate}. The numbers are based on the projections contained in the ITRS update of 2006, and are premised on the transition from the conventional bulk device to an ultra thin-body fully depleted device at the 32nm technology node. This is the reason for the nonmonotonic trends in V_{th} and $3\sigma_{V_{th}}$ observed in the figure.

2.5 THIN FILM THICKNESS

The thickness of the dielectric film that isolates the gate from the silicon channel greatly influences the transistor's electrical properties, including current drive, threshold voltage, and leakage current. Silicon dioxide (oxide) has been traditionally used as the gate isolation material. The scaling in oxide thickness has continued at the typical rate of a 30% reduction per technology generation and is currently approaching 10–12Å. The continued scaling of the oxide is, however, threatened as the current values of oxide thickness are approaching the physical limit of film scaling. The primary reason is the quantum-mechanical electron tunneling through the isolating dielectric material. Around the 65nm technology node, the gate tunneling current will become comparable to or greater than the channel leakage current. In one example [37], a 100nm process with T_{ox} = 16Å has the channel leakage of 0.3nA/μm of gate width, while the gate current is 0.65nA/µm.

The problem is especially severe for NMOS devices. PMOS devices also exhibit gate tunneling current, but for the same physical T_{ox}, it is typically an order of magnitude smaller that that of NMOS devices. The reason is

that holes have a higher effective mass than electrons and their tunneling probability is thus much smaller. This ratio is dependent on the material, however. For some alternative dielectrics, for example, nitrided oxides, the hole tunneling can become equal to electron tunneling [38].

The gate tunneling current through the currently used oxide (with thickness of 8–12Å) is so large, that no further reduction is possible. New dielectric materials with a higher value of the dielectric constant are sought to replace oxide. Several alternative materials with a range of dielectric constants have been explored. Some materials promise a great increase of the dielectric constant to the range of 25–50, compared to 3.9 for SiO_2, such as hafnium oxide (HfO_2). If successful, these materials will alleviate the problem of gate leakage. However, the quality of the insulator-silicon interface remains a problem, and major integration difficulties have been encountered for many such materials. The most realistic short-term hope comes from a nitrided gate oxide (oxynitride) with a dielectric constant in the range of 4.1–4.2. While providing less spectacular benefits, this material still leads to a $10\times$ reduction in gate leakage current [37].

Silicon dioxide films are created with a thermal oxidation process which historically has been extremely tightly controlled. The 3σ variation of oxide thickness has been around 4% [19]. Currently, the thickness of the oxide layer has reached a scale of atomic level roughness of the oxide-silicon interface layer [14], [10]. The Si-SiO_2 interface has a standard deviation on the order of 2Å [39]. The thickness of oxide film of 10Å corresponds to approximately five atomic layers of SiO_2, while the thickness variation is 1–2 atomic layers. Thus, the control of the interface layer and the oxide layer itself has become increasingly difficult, and is now governed by the fundamental limitations of interface roughness and atomic scale discreteness. That leads to growing variations in electrostatic device characteristics such as mobility and threshold voltage [14]. Most significant is its impact on gate tunneling current. Gate tunneling current shows an extraordinarily high sensitivity to the dielectic thickness [37]. For a device with $T_{ox} = 15$Å, and $\sigma_{T_{ox}} = 1.8$Å, the current can be $5\times$ larger than at the nominal conditions [40].

The variance of the T_{ox} variation is not a sufficient metric for analyzing the impact of oxide thickness variation on the electrical device properties. This is due to the need to consider the frequency distribution of the variation profile and take into account the correlation distance. A silicon-oxide interface is typically represented by a 2D Gaussian, or exponential, autocorrelation function with a given correlation length and the magnitude of variance [14]. Data shows that depending on the atomic orientation of the silicon substrate lattice, the interface roughness steps are on the scale of 1–3Å. Because of the difficulty of accurately studying atomic–level interfaces, there is a fairly large range of correlation distance values that have been experimentally reported. TEM measurements typically indicate a correlation length of 1–3nm, while AFM measurements are in the 10–30nm range [14]. The currently accepted view is that the correlation length (as determined by fitting roughness data

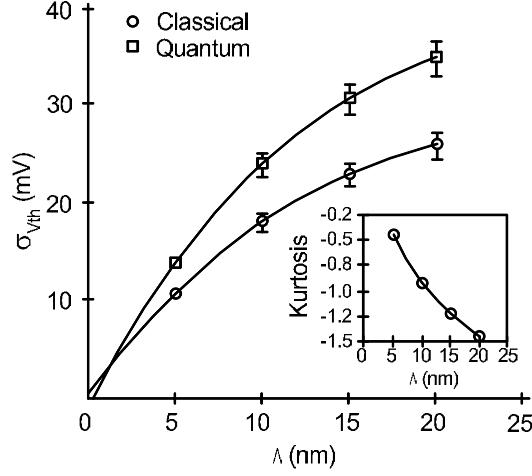


Figure 2.17 Threshold voltage uncertainty due to oxide thickness variation strongly depends on the correlation distance (Λ) characteristic of Si-SiO2 interface (Reprinted from [48], ©2003 IEEE)

to surface mobility data) is closer to the low range of the reported values, and the reasonable values to use are 7–15Å [47].

The impact of interface roughness and oxide layer nonuniformity on electrostatic device properties can be analyzed via careful 3D simulation [48]. For a device with an average T_{ox}=10.5Å, interface roughness steps of 3Å, and a correlation length of $\Lambda = 15$Å, it is found that the interface roughness leads to a V_{th} uncertainty of about $\sigma_{V_{th}} = 4$mV. Given the large range of reported values of correlation length, it is useful to study its impact on the threshold voltage uncertainty. Projected magnitudes of $\sigma_{V_{th}}$ based both on classical and quantum-mechanical simulations for several values of L_{eff} are presented in Figure 2.17 for a range of correlation length values. The assumed interface roughness value is 3Å, which is characteristic of the empirically measured devices. For the correlation length values at the higher end of the reported range (e.g. $\Lambda = 25$nm), the uncertainty in the threshold voltage is much larger: $\sigma_{V_{th}} = 35$mV, when quantum-mechanical effects are taken into account. The figure indicates that when the correlation length is much smaller than the characteristic MOSFET dimensions, the standard deviation of V_{th} depends linearly on the correlation length. For this linear range, the following model has been proposed to predict the geometry dependence of $\sigma_{V_{th}}$:

$$\sigma_{Vth} = \sigma_{Vth}^{\max} \Lambda / \sqrt{W_{eff} L_{eff}} \qquad (2.5)$$

where $\sigma_{Vth}^{\max} = 49$mV. The numerical simulations validate the above dependence of $\sigma_{V_{th}}$ on the FET dimensions. Overall, these results indicate that threshold voltage uncertainty due to oxide nonuniformity is, indeed, significant

when device dimensions are on the order of the interface correlation length. In devices below 30nm, this uncertainty is comparable to that introduced by random dopant fluctuations [48]. Experiments confirm that the two sources of V_{th} variability behave in an uncorrelated fashion. The total V_{th} variance is thus:

$$\sigma^2_{V_{th}} = (\sigma^{OTV}_{V_{th}})^2 + (\sigma^{RDF}_{V_{th}})^2 \tag{2.6}$$

Figure 2.17 also contains an inset showing the kurtosis of the V_{th} distribution as a function of the correlation length. Kurtosis is the measure of how non-Gaussian the distribution is. We see that for small values of Λ, the distribution is nearly Gaussian (small absolute value of kurtosis), but becomes increasingly flattened for larger correlation lengths.

2.6 LATTICE STRESS

A fairly recent systematic variability mechanism due to the impact of strain on device functionality has become increasingly important. One of the actively-pursued approaches to device engineering is the use of strained silicon to enhance circuit performance. The mobility is a strong function of stress: a physical stress on silicon lattice leads to the higher carrier mobility. This means that the transistor current drive and switching speed are also dependent on stress. The precise device physics of stress-induced mobility enhancement is quite complex. It is believed that strain enhances the electron mobility by reducing both the effective electron mass and the scattering rate. The hole mobility appears to be affected only by the effective mass change [41]. In addition to the above, stress appears to affect velocity saturation, threshold voltage, and current drive, with the effect on current drive (via mobility change) being the most influential. Stress in silicon can be created by adding layers of other materials that mechanically expand or compress bonds between the silicon atoms. The desired stress is tensile for NMOS and compressive for PMOS transistors. For creating strain in NMOS transistors a layer of silicon nitride is used, whereas PMOS transistors can be stressed by using silicon germanium. Electron mobility enhancements of up to 60% have been reported [42].

Importantly, stress can also be created as a by-product of the processing steps involved in traditional device fabrication. The cause of such stress is the mismatch in thermal expansion coefficients and oxidation volume expansion [43], [44]. The use of shallow trench isolation (STI) has been shown to lead to substantial compressive stress creation due to the above mechanisms; specifically the stress arises from the oxidation step that follows the formation of STI. NMOS mobility can be degraded by as much as 13% due to the stress caused by the proximity to the STI edge [41]. In addition to affecting mobility, the mechanical stress at the STI corners has also been implicated in anomalous leakage current.

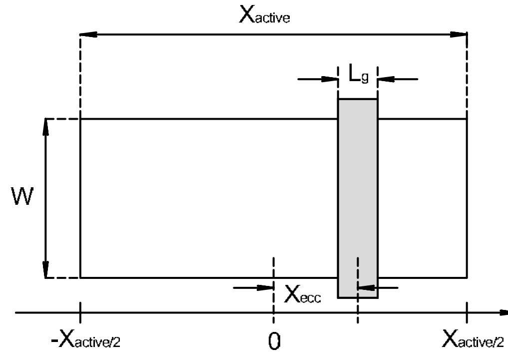


Figure 2.18 The stress is highly dependent on the size of active area (X_{active}) and the proximity of the poly to the edge (Reprinted from [45], ©2003 IEEE)

The STI-caused stress and its impact on mobility (and on-current) is highly dependent on the layout, specifically, the size of the active area and the distance to the STI edge. Because the stress produced by the STI is compressive, the trends are opposite for NMOS and PMOS devices: compressive stress enhances hole mobility and degrades electron mobility. For NMOS devices, the on-current is degraded as the active area is reduced. Consider the layout shown in Figure 2.18. For the length of the active area (X_{active}) below 5μm, the current drive is reduced up to 13% for a narrow-width device. At the same time, the PMOS current drive is increased by up to 10% for a narrow-width device, as in Figure 2.19. Additionally, the degradation (for NMOS) and enhancement (for PMOS) get bigger with the growing proximity to the STI edge. The dependence on the width of the poly-line is also quite significant, degrading both the NMOS (by 2%) and PMOS currents (by 10%) [45].

Thus, as the transistor active area shrinks and the channel is placed closer to the STI (trench) edge, the mobility degradation can be expected to become more significant. From the design point of view it is important that the amount of stress and therefore the electrical device characteristics are highly systematic and depend on the layout. As a result, transistors laid out with relatively wide spacing will perform quite differently from transistors laid out with high density for the same polysilicon dimensions.

2.7 VARIABILITY IN EMERGING DEVICES

In response to multiple challenges in device engineering, novel device architectures have been explored. The primary driver behind the search for alternative device architectures is the need to counteract the severe short-channel effects of bulk FETs and partially depleted SOI devices. New materials are also used in addition to novel device architectures to increase transistor performance and current drive; most notably, by increasing mobility of electrons

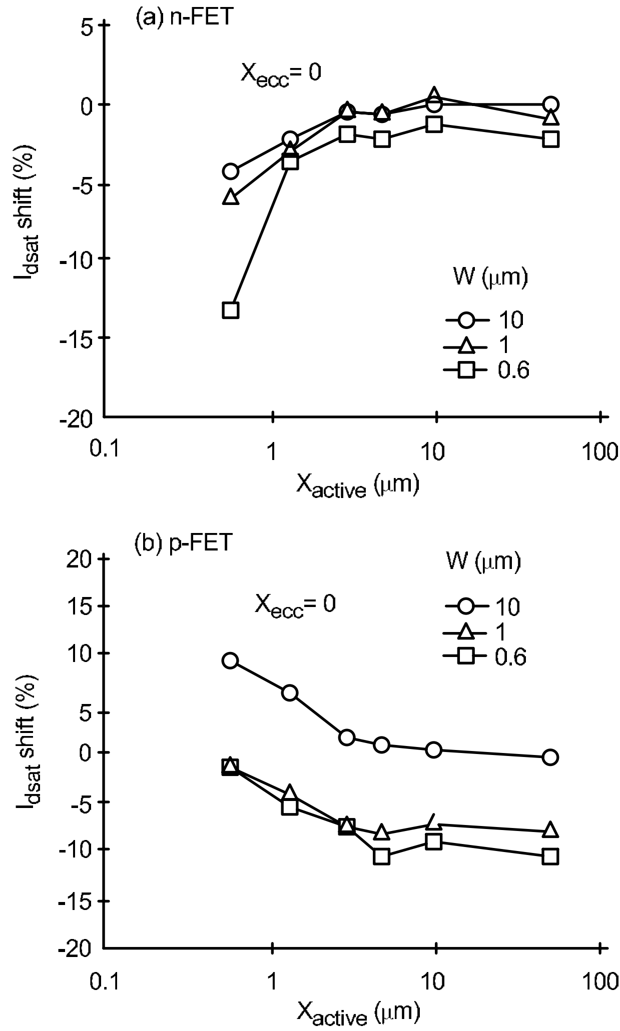


Figure 2.19 Impact on mobility is dependent on the size of active area. The trends are opposite for NMOS and PMOS devices (Reprinted from [45], ©2003 IEEE)

and holes. This mobility enhancement is achieved by introducing intentional stress into silicon lattice. Tensile strain increases electron mobility while compressive strain enhances the mobility of holes. However, in a way similar to the just considered unintentional strain due to trench fill in STI, devices that use strained silicon exhibit strong dependence of their transport properties on the layout specifics [50]. Experiments show substantial, on the order of 10–15%, dependence of carrier mobility on layout attributes, such as gate-to-gate spacing, length of the source and drain regions, and active area size.

The new device architectures aim at reducing the severity of threshold voltage roll-off and drain-induced barrier lowering. These device architectures include fully depleted silicon-on-insulator devices (FDSOI), dual-gate devices (e.g., FinFET), Tri-Gate, and Back-Gate devices. One common characteristic of these new device architectures is that they have thin fully depleted silicon body. This leads to two implications, important from the point of view of variability. First, because the channel is fully depleted, the device threshold voltage exhibits a stronger linear dependence on the doping concentration, compared to the power of 0.4 dependence in bulk FETs [49]. As a result, the variation in V_{th} due to random dopant fluctuation is more significant. Secondly, the thickness of the silicon body now has an influence on V_{th}, and thus variability in body thickness contributes to the variability in V_{th}.

One of the most interesting practical alternatives to the traditional planar MOS transistor is a dual-gate transistor, such as FinFET [46]. Planar MOSFET has one-sided control over the channel and has high leakage. A dual gate MOSFET has more electrostatic control over the channel, and thus has less leakage. The variation of V_{th} is in fact due to several distinct physical causes, including the short-channel effect, V_{th} dependence on the thickness of the silicon channel (fin), and the uncertainty due to random dopant fluctuations [36]. In a FinFET, the gate surrounds the thin silicon block (i.e., fin), forming the conducting channel on both sides of the fin. The threshold voltage of the FinFET strongly depends on the thickness of the silicon fin, as shown in Figure 2.20. The most severe variability issue in such devices is likely to be the channel thickness control. In the case of the vertical channel, its thickness is defined by a lateral lithographic process, and its tolerance is usually worse than

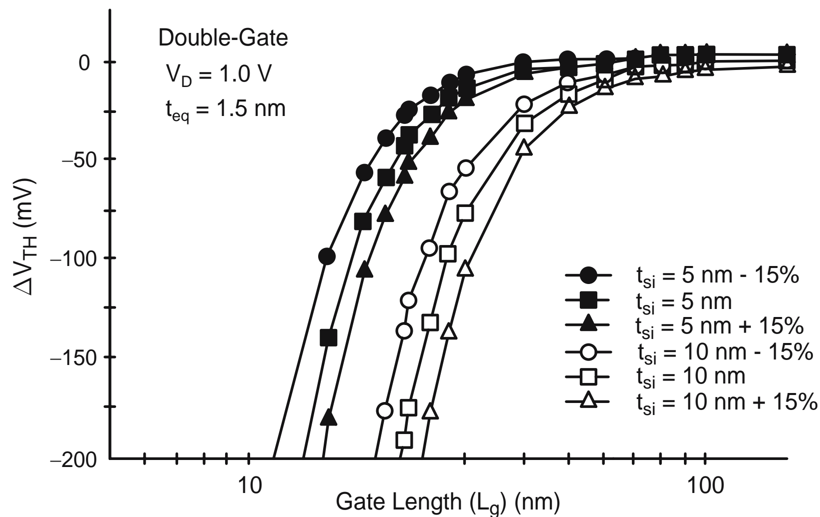


Figure 2.20 The amount of V_{th} variability in double-gate devices will be a significant concern (Reprinted from [11], ©1999 IEEE)

the film deposition (thermal growth) step used for the classical (planar) devices. The control of Si fin thickness therefore directly determines the degree of control on threshold voltage. Evaluation of all these factors indicates that the standard deviation of V_{th} for FinFETs with fin-thickness of 5nm will be about 100mV, of which only 25–50mV is due to random dopant fluctuation [36]. Because of the vertical channel in FinFETs, the transistor width is quantized to the number of silicon fins. The vertical variations in the fin height manifest themselves as FET width variations. It is interesting to observe that in this case, the global variations in fin height will lead to the same relative variation device widths, regardless of the absolute value of transistor width [49].

2.8 PHYSICAL VARIATIONS DUE TO AGING AND WEAROUT

In this book we are primarily concerned with uncertainty in the physical parameters of ICs resulting from the manufacturing process or the intrinsic device uncertainty. A different type of uncertainty that affects the physical parameters is caused by *temporal* factors.

The impact of the above physical mechanisms is to change the properties of devices over time. The difference is that the manufacturing and intrinsic variations are manifest at time zero, while the “temporal” variations appear over time. From the designer’s point of view, the impact of these changes is not different from the variability induced by the manufacturing process: the impact of both is to introduce uncertainty about the device properties. The traditional approach designers use to deal with these two types of variability mechanisms is also the same — to use margins and worst-casing. To account for the temporal effects, device models containing “aged” parameters are created to ensure that the circuit will operate under end-of-life conditions.

One useful way of comprehensively describing all sources of variability is by identifying their time constants. Depending on the time constant associated with the mechanism of variability, it is useful to divide them into two groups. The fast, small time constant temporal variability mechanisms include effects such as SOI history effect and self-heating. The second group of mechanisms has a much longer time constant and is related to aging and wear-out in physical parameters of transistors and interconnects. The primary mechanisms in this category include: (i) negative-bias temperature instability, (ii) hot carrier effects, (iii) electromigration.

Negative bias temperature instability (NBTI) affects p-channel MOSFETs. Its impact is to increase over time the threshold voltage of the p-FET, which reduces its current drive capability and thus increases circuit speed. At some point, the possibility of path timing violations arises. NBTI is due to the creation of interface traps and the positive trapped charge. The NBTI stress occurs when the p-FET is on with gate voltage $V_g = 0$ and $V_d = V_s = V_{dd}$. When stressed continuously for the course of the device lifetime (e.g., 10 years) the

p-FET threshold voltage can change by as much as 42%. Empirical observations show, however, that when the stress is removed, the NBTI can be reversed to some extent, even if not entirely. Since in a real circuit environment transistors typically are not stressed continuously, the true NBTI lifetime can be much longer. Alternatively, the increase of the threshold voltage is much smaller over the same period of time. For devices in 65nm technology, the lifetime computation that takes into account the real dynamic of the device switching predicts a V_{th} degradation of 38%. While the threshold voltage change is only 5% less severe than under static conditions, the lifetime is effectively doubled, since in the dynamic case it will take 10 more years (20 years in total) to experience the same level of degradation as in the static case.

In addition to the stress patterns that are determined by the workload (e.g., the activity factor), the amount of threshold voltage degradation due to NBTI depends on the supply voltage and temperature in the device vicinity, as well as the capacitive load driven by a gate and the gate design factors (e.g., the ratio of p-FET and n-FET device geometries) [49].

Hot carrier effect (HCE) affects primarily n-channel MOSFETs. It is due to the injection of additional electrons into the gate oxide near the interface with silicon. During switching, the electrons gain high kinetic energy under the influence of the high electric field in the channel. Depending on the relative voltages on the FET terminals, different mechanisms may be responsible for electron injection into the oxide, including (i) the direct channel hot electron injection, (ii) the secondary generated hot electron injection, (iii) the drain avalanche hot carrier injection, and (iv) the substrate hot electron injection. Regardless of the specific mechanism of injection, the ultimate result is the growing interface charge that leads to the increase of the threshold voltage, lower current drive, and longer switching time. The danger of HCE is that the timing constraints of some paths may be violated at some point. To prevent this from happening, the design must be checked with the aged models that correspond to the end-of-life value of the threshold voltage.

Electromigration is the process that affects wires and is caused by the continuous impact of high current densities on the atomic structure of the wire. Under the influence of current flow, the atoms of the metal wire may be dislocated. This may ultimately lead to the creation of shorts between the wires when the dislocated atoms of two neighboring wires are contacted. This may also lead to the creation of an open failure in the wire when the dislocated atoms produce a void in the wire that damages its electrical connectivity.

2.9 SUMMARY

Variability in the front-end of the process technology will continue to be the main contributor to the overall budget of variability. There are multiple systematic design-process dependencies (proximity, etch, stress) that are of

first-rate importance. Because of their systematic nature, their impact on design can be eased by improving the characterization and modeling of these effects, and propagating the appropriate information to the designer. The fundamental, or intrinsic, components of variation are essentially random. Their impact on the design will continue to grow, requiring a substantial change in the design approaches and practices.