Device and Circuit Techniques for Reducing Variation in Nanoscale SRAM

Andrew Evert Carlson

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-51

http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-51.html

May 15, 2008

Acknowledgement

The author gratefully acknowledges the support of Profs. Tsu-Jae King Liu and Borivoje Nikolic, Dr. Srinath Krishnan, the UC Berkeley Device Group students, and the staff of the Microlab. The author is also grateful for financial support from AMD and the SRC and silicon support from AMD, ST Microelectronics, and the UC Berkeley Microlab.

Device and Circuit Techniques for Reducing Variation in Nanoscale SRAM

by

Andrew Evert Carlson

S.B. (Harvard University) 2003
M.S. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy
in

Engineering - Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Borivoje Nikolic
Professor Robert Leachman

Spring 2008

Device and Circuit Techniques for Reducing Variation in Nanoscale SRAM

by

Andrew Evert Carlson

Abstract

Device and Circuit Techniques for Reducing Variation in Nanoscale SRAM

by

Andrew Evert Carlson

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

SRAM scaling, a major driver of microprocessor development, is threatened by increasing variation in transistor parameters such as threshold voltage and gate length. With a target-based model for device I-V characteristics, the effects of these variations on SRAM performance can be well understood and predicted. A robust, iterative algorithm for estimating SRAM cell yield is developed. The analysis is extended to time-dependent reliability problems, and a statistical methodology for robust cell design is presented.

For future technology nodes, SRAM scaling will require device and circuit innovations to suppress variation. Multi-gate devices and extended spacer lithography processes can be used to reduce random variability at its source. Feedback circuits can be used to reduce systematic SRAM variation after fabrication. Implementation of any one of these techniques is expected to result in a significant yield improvement of several sigma. In combination, these techniques are expected to enable robust SRAM scaling to the end of the roadmap.

Professor Tsu-Jae King Liu
Dissertation Committee Chair

To Christina, my light at the end of the tunnel

Contents

List of Figures

List of Tables

Acknowledgements

It is not so much that a Ph. D. is a lot of work (though it is) as it is a lot of time. During my time at Berkeley, several people have guided my work, shaped my career goals, and helped my education. I am grateful for all of their help.

I would like to begin by acknowledging the steadfast support of my advisor, Prof. Tsu-Jae King Liu, who has always been generous in allowing me to pursue my own research interests with my own direction, all the while providing technical guidance, encouragement, and financial support. Research advisors teach the most by example, and there is much about her management style and professional conduct that I hope to bring with me as I go into industry.

I am deeply grateful for all the help of Prof. Borivoje Nikolić, who has been tremendously gracious in unofficially advising my research even though I was not his student. His perspective on SRAM and guidance toward opportunities with a meaningful contribution were exceptionally helpful. Although he was hard for me to read at first, I became profoundly impressed with the sense of fairness and morality he brings to his work.

Dr. Srinath Krishnan provided invaluable opportunities and guidance while and since my internship at AMD, for which I am also very grateful. He encouraged me to develop an SRAM yield model, which has provided a great advantage toward understanding SRAM and all its tradeoffs. I hope to be as informative and impactful with my coworkers as he is with his.

There are several current and former students and staff whom I would like to thank with regards to this work specifically. With Zheng Guo and Liang-Teck Pang I had several helpful SRAM discussions. They also provided invaluable assistance with the logistics of tapeout and testing. Dr. Sriram Balasubramanian was an astute and friendly opponent in many informal debates, who helped me sharpen my technical arguments while he was a student here. Xin Sun and Changhwan Shin provided simulation data for the triple-gate bulk devices and helped me stay sharp on device theory. Xin, Dr. Vidya Varadarajan, Joanna Lai, and especially Albert Lai helped me debug different processes in the UC Berkeley Microlab. Evan Stateler and Jay Morford also helped me in the lab to get the machines working.

In addition to the above, several others have provided indirect assistance, by helping with my education here at Berkeley. Hideki Takeuchi taught me, among many other things, the technical stubbornness needed to get things done in the lab. Pankaj Kalra provided a much-needed baseline for the Ph. D. experience because he shared my perspective. Donovan Lee, Steve Volkman, Dr. Alvaro Padilla, Dr. Kyoungsub Shin, Dr. Dan Good, Alejandro de la Fuente Vornbrock, Hei Kam, Kinyip Phoa, Yu-Chih Tseng, Noel Arellano, and Prof. Nathan Cheung have also provided technical or educational assistance in some form. Thank you to all.

This work has been funded through several sources. I am first thankful to John Ennals and Bill En of AMD for their help in winning the Semiconductor Research Corporation (SRC)/AMD Mahboob Khan Fellowship. This work has also been supported by the Center for Circuit and System Solutions (C2S2) Focus Center, one of five research centers funded under the Focus Center Research Program, an SRC program. Spacer lithography processing was completed in the UC Berkeley Microlab. Fabrication of a 45nm chip was donated by STMicroelectronics. Data for 65nm and 90nm silicon was provided by AMD.

Finally, I thankfully acknowledge the support of my fiancée, Christina, who with her compassion and understanding helped me through the darkest times, even when she was far away.

“Many complain about their memory, few about their logic.”

— Adapted from Benjamin Franklin

Chapter 1

Introduction: SRAM Scaling

SRAM scaling represents one of the greatest challenges to decreasing cost per function in microprocessors. On-chip cache size has become increasingly important for high performance applications, and it now presents more of a limit to microprocessor speed than clock rate. The models and methodologies for the design of SRAM, an integral component of microprocessor cache, have changed with time, as new tradeoffs and constraints have emerged. Currently, continued scaling is threatened by variability in SRAM performance and function. This work addresses the emerging threat of variability in three ways: by advancing the understanding of the mechanisms of variation-induced SRAM failure, by developing new devices and processes to address the sources of variation, and by proposing new circuit techniques to compensate for existing variation.

1.1 Static Random Access Memory

1.1.1 Cell Architectures

An SRAM array is composed of many identical cells, small circuits that can each store a single bit of information. The most common type of cell, the 6-T SRAM (Fig. 1.1a), is named for the six transistors which comprise it. The cell consists of two cross-coupled

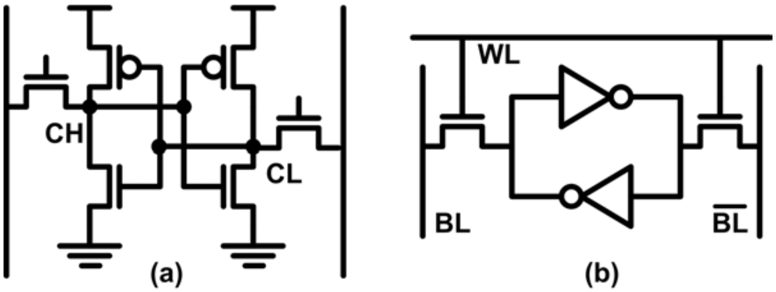

Figure 1.1. The most common SRAM cell architecture, the 6-T SRAM has two pull-up transistors, two-pull-down transistors, and two pass-gate transistors (a). The pull-up and pull-down transistors make up two cross-coupled inverters (b). Cells are accessed by means of orthogonally-routed wordlines, **WL**, and bitlines, **BL** and $\overline{\textbf{BL}}$.

inverters (Fig. 1.1b), made up of the PMOS pull-up devices and the NMOS pull-down devices. The cross-coupled inverters ensure that the internal nodes of the cell always contain complementary values. Two NMOS pass-gate devices connect the internal nodes of the cell to array-level bitlines and provide read and write access to the cell.

The 6-T SRAM is operated in the following way. To read the cell, the bitlines are precharged to a high bias and the wordline voltage is raised. On the side of the cell storing the logical zero, the bitline is discharged through the access transistor. Depending whether the bitline on the "cell high" (**CH**) or "cell low" (**CL**) side is discharged, the cell is read as a logical one or logical zero. To write the cell, the bitlines are driven to complementary values and the wordline voltage is raised. On the side of the cell with the bitline at a low bias, the internal node is discharged through the pass-gate transistor. The cross-coupled inverters raise the bias on the opposite node and latch the new voltages in place.

From this simple description the two basic modes of SRAM failure can be understood. During a read operation, the bias of the low internal node will increase, due to the current through the pass-gate. If the bias rises above the switching point of the inverters, the cell becomes unstable and may switch its state. This event is called a read disturb (Fig. 1.2a). It can be prevented by ensuring the pull-down transistors are much stronger than the pass-gate transistors. This ratio of device strengths is called the *beta ratio* and was an important parameter in early SRAM design.

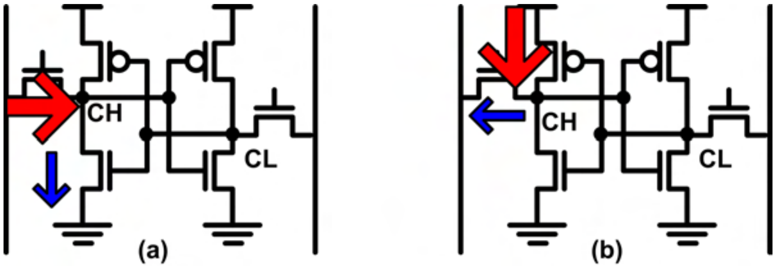

Figure 1.2. The two basic modes of SRAM failure are a read disturb (a), *e.g.* in which discharge current through the pass-gate device raises V_{CH} from a logical zero to a logical one, and a write fail (b), *e.g.* in which discharge current through the pass-gate is unable to lower V_{CH} from a logical one to a logical zero.

The other primary mode of SRAM failure is a write failure (Fig. 1.2b). During a write operation, the discharge of the internal node through the pass-gate must overcome a restorative pull-up current through the PMOS device. Write failures can be prevented by ensuring the pass-gate transistors are much stronger than the pull-up transistors. The *gamma ratio* measures this quantity.

The 6-T cell remains the favorite of SRAM architectures because of these two simple tradeoffs. A design can be virtually guaranteed to work by sizing the devices for high beta and gamma ratios, but at the expense of cell area. In spite of the additional device and processing challenges present in modern SRAM design, these fundamental tradeoffs with area still hold true today.

To improve this tradeoff, several alternative SRAM architectures have been investigated. The 4-T SRAM (Fig. 1.3a) removes two transistors from the inverters of the 6-T design [1, 2]. 4-T SRAM has been shown to exhibit better read stability than 6-T for high supply voltages [3] and for low voltages with independently-gated double-gate transistors [4, 5], such as FinFETs [6]. The internal nodes hold complementary values as in the 6-T design; however, the charge on the high bias node is supplied only during the write or through delicate balancing of device off-state currents. It is therefore vulnerable to discharge during a read operation or through leakage paths in the cell and requires periodic refreshing. It is also susceptible to variability [5]. The recently-proposed 8-T SRAM (Fig. 1.3b) adds two transistors to the 6-T cell as a separate read port [7]. It enhances read stability by

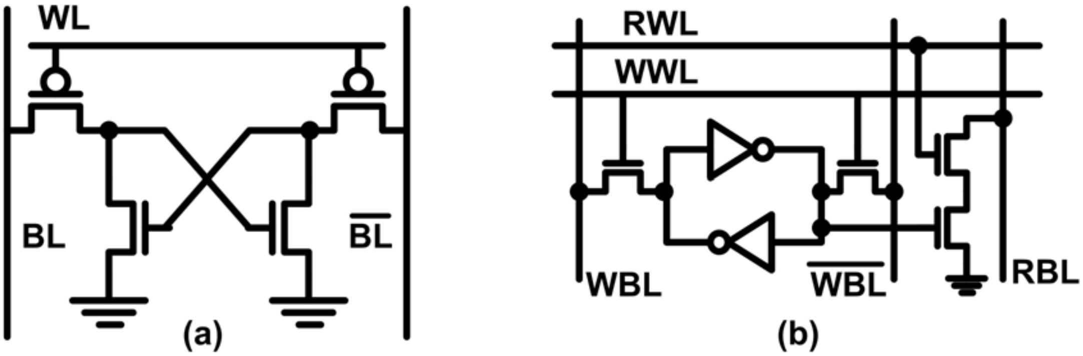

Figure 1.3. The 4-T SRAM has a smaller area but is less robust than the 6-T cell [1, 2] (a). The 8-T SRAM decouples the read and write operations to allow for simultaneous enhancement, but it has a larger cell area and requires separate read and write bitlines and wordlines [7] (b).

eliminating bitline discharge into the internal node, but at the expense of a larger cell. The increase in cell area can be reduced by improvements in array efficiency through specific addressing schemes, but not completely [8]. It is not yet clear how the write yield compares to that of a 6-T design of comparable cell area and array size. Other SRAM architectures, including 9-T [9] and 10-T [10, 11] have been proposed, but also have undesirable tradeoffs in area, reliability, or performance compared to the 6-T design. The analysis presented in this work therefore assumes a 6-T cell; however, the models and methods developed could be easily extended to other cell architectures.

1.1.2 The Drive to Scale

In the design of SRAM, cell area is invariably the metric to optimize. Other metrics, such as read stability or access times, are important insofar as constraints are met, but smaller area is always the primary goal. SRAM scaling has historically followed Moore's Law, with the same economic drivers of speed and cost per function (Fig. 1.4).

SRAM scaling reduces memory access times by allowing more memory to be closer to a logic core. In early microprocessors, before logic and memory were integrated on the same chip, motherboard-based RAM arrays were used as cache memory to reduce the delays associated with hard disk access. Tiny clusters of RAM cells were used as registers

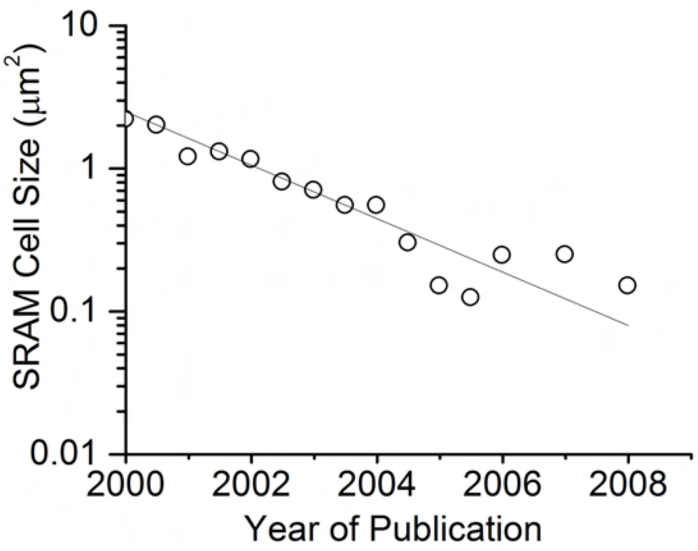

Figure 1.4. Reported SRAM sizes have historically followed Moore's Law, with an area reduction of 0.5× every 18 months. Data from [12, 7, 13, 14, 15].

within the core to accelerate program execution further. As transistor dimensions scaled, it became possible to embed a memory cache in the same chip as the logic core, eliminating the I/O and parasitic delays associated with off-chip communication, and memory access times were drastically reduced. In fact, the development of embedded SRAM was instrumental in defining a niche for SRAM among other types of memory, such as dynamic random access memory (DRAM), non-volatile EEPROMs, and magnetic disks. Cache evolved into a multi-tiered structure, with SRAM making up the fastest access memory (Fig. 1.5a). In modern microprocessors, this trend has continued, such that SRAM cache itself has multiple levels. For example, in recent Intel and AMD microprocessors, L1 (level one) cache contains a small amount of the fastest memory cells. At the next level, L2 cache contains a large amount of cells that are slightly slower, and so on. Memory access time is an emerging constraint on the performance of microprocessors, and is best reduced by increasing the SRAM cache size. Presently, a microprocessor's L2 cache size is a commonly quoted specification, second only to clock speed. Scaling cell area allows for a larger cache in the same area. Yamagata notes that transistor scaling has not led to a commensurate reduction in microprocessor die

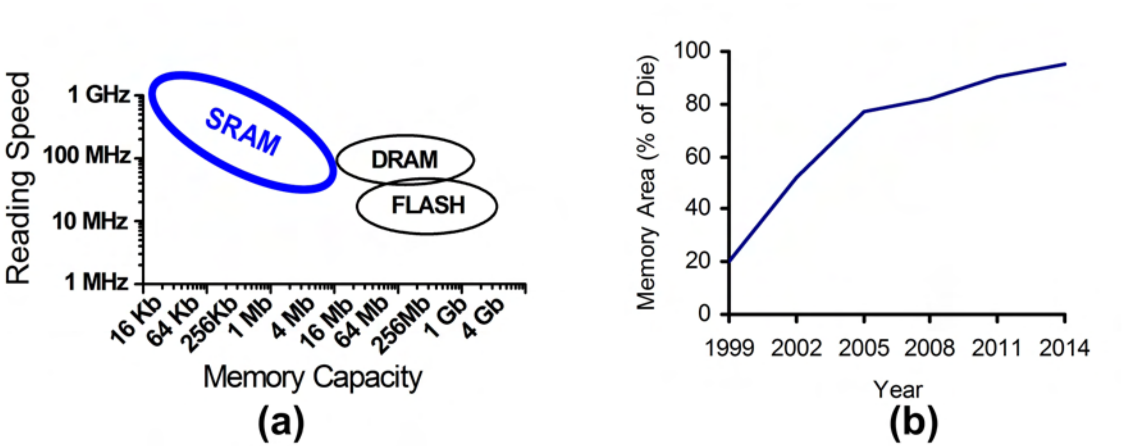

Figure 1.5. SRAM is used in relatively small, on-chip cache memories for fastest access (a). With continued scaling, the physical size of a microprocessor remains approximately the same, and the additional area is filled with more memory (b). Figures adapted from [16].

size, but rather more functionality has been integrated into designs of a comparable area [16]. With ever increasing proportion, more functionality means more memory (Fig. 1.5b).

Together with the aforementioned tradeoff between cell area and functionality, the drive to scale resulted in cell designs meeting minimum constraints in read stability. Early SRAM designs could make use of minimum-width devices for the pull-up and pass-gate devices. The gamma ratio for such a device would be a function of the mobilities in the process, $\gamma \approx \mu_n/\mu_p$, which ensured sufficient write-ability before the advent of strained silicon technology. The pull-down devices would be sized larger to meet the minimum beta ratio needed to ensure stability. Such a cell design had the benefit of being directly applicable to a new technology node with a simple shrink. A standard dimension reduction of 0.7x results in a cell area of 0.5x. SRAM scaling was therefore automatic with transistor scaling, and advanced models and custom design rules were not necessary.

In fact, the only metric of significant concern in early SRAM was that of read stability, since the minimum beta ratio was desired to reduce cell area. Since Seevinck's seminal work in 1987, read stability has been quantified with the static noise margin (SNM), which is defined as the minimum amount of noise needed to upset the state of the cell [17]. It is

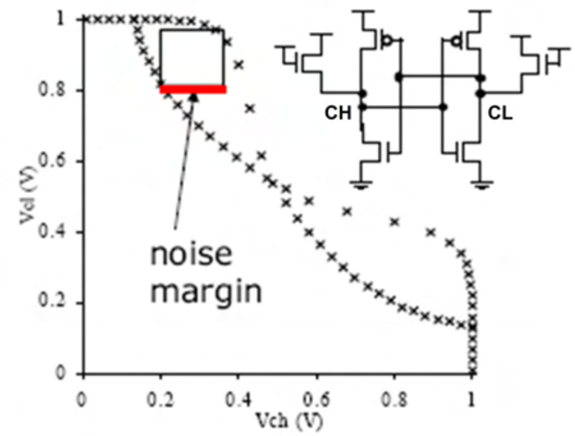

Figure 1.6. Static Noise Margin (SNM), a metric for read stability, can be illustrated with the cell's voltage transfer characteristics (sometimes called the butterfly curves). The curves are generated by sweeping the voltage of one internal node and measuring the voltage of the opposite node with the cell biased as shown. SNM is represented by the side of the largest square that fits within the curves. These curves were generated from measurements of a fabricated cell in an industrial 90nm SOI process.

commonly illustrated with the voltage transfer characteristics (Fig. 1.6, sometimes called the butterfly curves) for the SRAM cell, in which it corresponds to the size of the largest square that fits within the curves. Cells with SNM values of at least 25% of the cell supply voltage, V_{DD}, are generally considered to have excellent read stability. High SNM cells generally feature a switching voltage near $V_{DD}/2$ and high inverter gains around this point. SNM remains the most significant SRAM metric today, but it is no longer sufficient to guarantee array functionality.

1.2 Scaling Issues for Embedded SRAM

In the past few years, SRAM scaling has faced increasing challenges. Short channel effects and the abandonment of constant field device scaling made existing SNM models obsolete. The development of strained channels, which improve PMOS mobility more than that of NMOS, has decreased cell write-ability to the point where its tradeoff with SNM has become a significant aspect of SRAM design. As dimensions shrink, variations in transistor performance degrade functionality and reduce yield. Devices which leak more in the off-state limit performance, constrain array architecture, and in extreme cases can cause cell instability. With less capacitance, the internal nodes of a scaled cell are more susceptible to leakage currents and other noise sources.

Transistor scaling

Under constant field scaling, all dimensions and voltages for a transistor were scaled down so that the electric fields remained constant between technology nodes. This not only maintained the same beta and gamma ratios for the scaled cell, it allowed SNM to be modeled with closed-form equations [17]. Subthreshold current could be ignored, and long channel equations for drain current (I_{DS}) could satisfactorily model transistor behavior:

$$I_{DS} = \begin{cases} \mu C_{ox}\frac{W}{2L}(V_{GS}-V_T)^2 & V_{GS} > V_T \text{ and } V_{DS} \geq V_{GS}-V_T \\ \mu C_{ox}\frac{W}{L}V_{DS}(V_{GS}-V_T-\frac{V_{DS}}{2}) & V_{GS} > V_T \text{ and } V_{DS} < V_{GS}-V_T \\ 0 & V_{GS} \leq V_T \end{cases} \tag{1.1}$$

where μ is the carrier mobility, C_{ox} is the gate oxide capacitance per unit area, W is the width of the device, L is the gate length, V_{GS} is the voltage on the gate with respect to the source, V_{DS} is the voltage on the drain with respect to the source, and V_T is the threshold voltage of the transistor.

In modern MOSFETs, subthreshold and gate leakage currents inhibit continued scaling of V_T and C_{ox}. Weak-inversion currents have become significant in determining the voltage transfer characteristics for modeling SNM. Furthermore, short channel effects such as drain-induced barrier lowering (DIBL), channel length modulation, and velocity saturation complicate the equations and make the old models obsolete. Of these effects, DIBL is particularly deleterious to SNM, since it reduces inverter gain at high supply voltages.

In addition, the ratio between on-currents for NMOS and PMOS has decreased with scaling, due in large part to the development of strained silicon channels (Fig. 1.7). Strain technologies have improved hole mobility μ_p more than electron mobility μ_n. Although beneficial for speed in logic devices, this can degrade SRAM write-ability. If minimum-

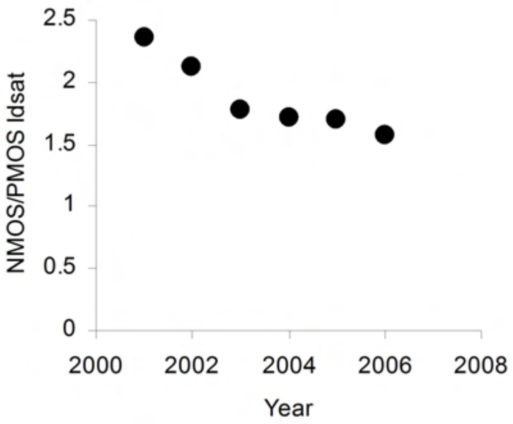

Figure 1.7. Recent advances in high performance CMOS have narrowed the gap between reported NMOS and PMOS drive currents (I_{dsat}), measured at $V_{DD} = 1.0V$ and $I_{off} = 100nA/\mu m$. [18, 19, 20, 21, 22, 14] For SRAM cells with minimum-width pull-up and pass-gate devices, this scaling trend results in a decrease in write-ability.

width pull-up and pass-gate devices are retained, the gamma ratio of the cell is reduced. The pass-gate must be made larger to maintain the original gamma ratio, but this requires a proportional increase to the pull-down device to maintain the same beta ratio. It thus becomes more difficult to scale cell area at the historical rate.

Variation

The problems caused by transistor scaling are exacerbated by the emergence of process variations. Variations in transistor parameters such as threshold voltage, gate length, or channel width affect the transistor's drive strength. In an SRAM cell, this may affect the SNM, write-ability, or access times. Symmetric circuits like the 6-T SRAM cell are especially vulnerable to mismatches in the strengths of paired transistors. As transistor dimensions scale down, the impact of process variations increases, and the cell yield drops.

The issue is compounded by the increasing SRAM array size. Cache sizes of several tens of million identical cells are common. To achieve high yield for the entire array, the nominal cell design must now have a very large margin for variation of at least five or six standard deviations. As cache sizes increase, the required margin will continue to grow.

Thus with continued SRAM scaling, cell yield will decrease even as arrays require higher yielding cells. This makes variation the greatest challenge to SRAM scaling.

Leakage current

In allowing greater subthreshold and gate leakage currents, transistor scaling can curb further SRAM scaling and degrade cell stability. Subthreshold leakage through the pass-gate transistors of many inactive cells can compete with the current through a single active cell to impair read access times. A constraint on the array column height, the number of cells on each bitline, may be needed to meet access time constraints [23]. Leakage currents through the power supply for several million cells can consume a significant portion of the power budget of a chip [24]. Gate leakage currents within the SRAM cell have been shown to degrade SNM and may also affect write-ability [25, 26]. With continued scaling, the capacitance on the internal nodes of a cell decreases. Thus the amount of charge needed to disturb a cell decreases, while the magnitude of leakage current increases.

Soft error rates

The reduction in capacitance is also significant for soft error events, in which the state of an SRAM cell is upset by the introduction of a large impulse of noise to the internal nodes. Soft error rates describe the frequency with which external events such as alpha particle collisions can cause a read disturb. As SRAM scales down, the incidence of soft error rates increases and poses a significant reliability challenge [27, 28].

In summary, there are several major challenges for continued SRAM scaling, and they are all growing worse. Scaled devices obsolete SRAM models and require new cell designs for each technology node. They are more sensitive to process variations, have increased leakage currents, and are more susceptible to external noise. Each of these issues is an area of current research. This work focuses on the most problematic of these, variation.

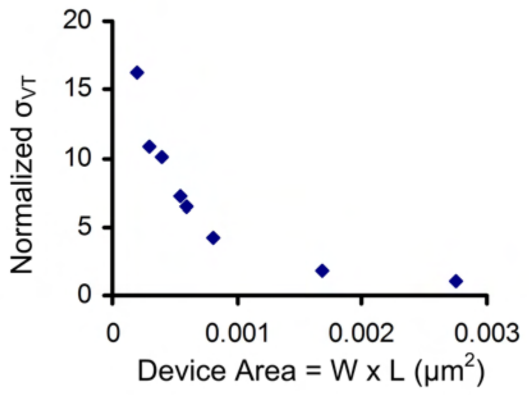

Figure 1.8. The variance of the threshold voltage of a MOSFET increases in inverse proportion to channel area due to random dopant fluctuation. The points in this plot are generated by Monte Carlo simulation, but the effect has been observed experimentally in several technologies [29].

1.3 Studies of Variation

1.3.1 Dopants and patterning

One of the most significant sources of process variations for current VLSI transistors ($L_G > 20$nm) is random dopant fluctuation [29]. To achieve a channel dopant concentration of 10^{19} atoms/cm^3 in a scaled MOSFET with dimensions less than 50nm, fewer than 100 dopant atoms are required. The displacement or absence of only a few dopants can result in threshold voltage variations. Fig. 1.8 illustrates the increase in the standard deviation of V_T as a function of channel area ($W \times L$) [29]. Threshold voltage variation due to random dopant fluctuation increases proportionally with $1/\sqrt{WL}$ [30]. With further scaling, discrete effects from displaced source and drain dopants may add to the variation. Recently, experimental studies have shown random dopant fluctuation is responsible for the majority of long channel V_T variation; however, it does not explain all the variation in NMOS V_T [31].

A second source of variation, which is becoming increasingly significant with continued scaling, is patterning. The edge of a printed line exhibits roughness on the scale of 5 nm, primarily due to polymerization effects in the photoresist [32]. This line edge roughness

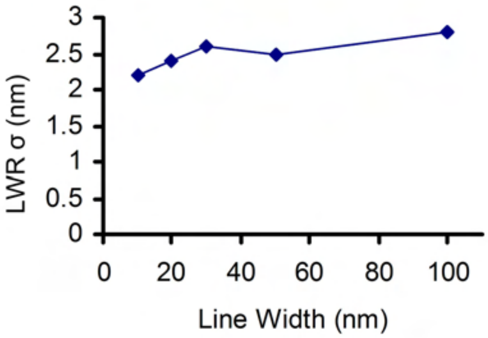

Figure 1.9. The variance of lithography-defined patterns becomes more significant with continued scaling, due to phenomena such as line edge roughness (LER) (figure adapted from [32]) and proximity effects [33, 34]. The effects are manifest in the gate lengths and channel widths of a $0.79\mu m^2$ SRAM cell after gate patterning [35].

(LER) becomes significant for dimensions smaller than 50 nm, such as gate length or channel width (Fig. 1.9). Although the variance in line width decreases as the nominal width scales down, proportionally its magnitude increases. This is especially significant for undoped multi-gate devices (e.g. FinFETs) in which V_T is set by the thickness of the active region. In addition to LER, a critical dimension can vary due to image effects from proximity or corner rounding. Printed patterns with sharp corners exhibit a rounding of the feature at spatial frequencies beyond the resolution of the lithography system, affecting the width of the feature near the corner. SRAM gate length has been shown to vary as a function of the layout of the gate and nearby features [33, 34].

Additional sources of variation can be present in strain application or contact resistance; however, these sources are not yet significant for SRAM.

1.3.2 Contemporary work

Variation in SRAM is currently an active area of research, with several yearly reports on measured yield or SNM specifically [36, 37, 38, 39, 21, 22, 40]. Among all types of variations, Venkatraman *et al.* reported that uncorrelated random variations dominate, based on measurements of 90nm node devices [41]. Yamaoka *et al.* reported that the standard deviation of these variations can depend on systematic variations at the array or wafer level [42]. In some designs, these variations can affect the write-ability of the cell more than SNM [43]. Statistical or "variation-aware" design methodologies are now indispensible [44, 42, 45].

Such methods require fast and accurate models for critical SRAM metrics. Unfortunately, the models with closed-form equations for SNM are no longer accurate for short-channel devices that operate near threshold. Recent models that do achieve a closed-form or semi-analytical expression for SNM make large approximations at the expense of accuracy [46, 47, 48]. They are fast, but not accurate. Other modeling efforts have derived new SRAM metrics [49, 50] or taken a probabilistic approach [51, 52], but lack the tractability or the fundamental basis of SNM. New metrics also take time to be embraced. Although several write-ability metrics have been proposed [49, 43, 53], a consensus has not yet emerged.

A common practice for SRAM modeling is circuit simulation, with a program such as SPICE, using advanced device models and Monte Carlo methods to estimate yield. This approach is accurate, but not fast. Accurate device models must be developed, which can be difficult and time-consuming in a developing technology. For transient simulations, the parasitic resistances and capacitances of a layout must also be modeled, which can require multiple iterations of process characterization. Monte Carlo simulations require many iterations as well. Nevertheless, this approach can yield useful evaluations of the sensitivities and variability of an SRAM design [54].

For process or device technologies where accurate circuit models are not yet available, the mixed-mode capability of a device-simulator such as TAURUS [55] is generally used. Although these simulations require even more computing time, they enable useful observations on the scaling behavior of SRAM. Such simulations have shown that multi-gate devices will be attractive for SRAM due to improved control of short channel effects and reduced variability [4, 56, 57, 58]. Furthermore, devices with undoped channels are expected to greatly reduce variability by mitigating random dopant fluctuation [59, 60].

Such simulations can influence the development of new devices to reduce variation. Dixit *et al.* have begun investigating the variability of fabricated FinFET SRAM using a spacer lithography patterning technique to reduce LER [61]. Okayama *et al.* introduced a fully silicided gate to reduce V_T variation from dopant penetration [62]. Reducing variation at the device level enables a higher-yielding SRAM cell.

In addition to these efforts, new circuits have been proposed to compensate for increasing variability. By modifying the body bias of blocks of several SRAM cells, yield metrics such as fail counts and minimum operating voltage (V_{min}) can be improved [63, 64]. Wordline biasing can also be used to tradeoff read stability and write-ability [65]. Read stability can be improved by limiting the amount of charge flowing into a cell [66, 67]. The primary tradeoff of these techniques is an array-level increase in area.

1.3.3 This work

This work aims to facilitate continued SRAM scaling in three ways: by furthering the understanding of variation and its sources in an SRAM cell and by developing a new modeling approach to accelerate statistical design methods, by investigating new devices and processes to reduce the sources of variation, and by proposing new SRAM circuits to compensate for increasing variability.

In chapter 2, a new modeling approach is presented that is both fast and accurate for read and write SRAM metrics. Unlike previous closed-form or semi-analytical models, this approach uses several device-specific I-V targets for improved accuracy. Approximations are made to the non-critical parts of the I-V curves, eliminating the need for time-consuming device simulation or model development. This modeling approach is used to investigate cell sensitivities. A statistical design methodology using these sensitivities is proposed as a fast alternative to Monte Carlo iteration. The model is used to provide insights into mechanisms of SRAM failure over time.

In chapter 3, methods to reduce process variation from random dopant fluctuation and lithography are proposed. Device architectures that do not rely exclusively on dopants to set V_T, such as undoped FinFETs, are proposed to enhance estimated SRAM yield. SRAM cells with straight active features are shown to have reduced variation due to lithography, and an extended spacer lithography process is developed to enable high-density integration with low variability. Spacer-defined circuit design is demonstrated for a 0.0512μm^2 SRAM cell, which could be scaled smaller than any previously-reported SRAM.

In chapter 4, circuit techniques to cope with process variation are presented. A circuit to sense and correct systematic and large-area variations is demonstrated to optimize the read / write tradeoff over a wide range of operating conditions. A technique to estimate process variability from SRAM metrics and probabilities is proposed, enabling SRAM measurements as a form of *in situ* characterization to accelerate process development. FinFET-based SRAM designs with independent gating are introduced and analyzed to enhance read stability and write-ability, allowing six sigma yield for supply voltages as low as 0.4V.

Individually or in combination, it is hoped that these techniques may advance SRAM development through the next several technology nodes. The modeling approach of chapter 2 is already starting to be adopted in industry. Some kind of transition to new devices, processes, or circuits is widely expected for SRAM specifically, and it is the goal of this work to help facilitate such a transition.

1.4 References

[1] R. F. Lyon and R. R. Schediwy. CMOS static memory with a new four transistor memory cell. *Proceeding of Stanford conference on advanced research in VLSI*, pages 111–131, 1987.

[2] K. Noda, K. Matsui, K. Imai, K. Inoue, K. Tokashiki, H. Kawamoto, K. Yoshida, K. Takeda, N. Nakamura, T. Kimura, H. Toyoshima, Y. Koishikawa, S. Maruyama, T. Saitoh, , and T. Tanigawa. A 1.9-μm2 loadless CMOS four-transistor SRAM cell in a 0.18-μm logic technology. *IEEE International Electron Devices Meeting*, pages 643–646, 1998.

[3] O. Semenov, A. Pavlov, and M. Sachdev. Sub-quarter micron SRAM cells stability in low-voltage operation: a comparative analysis. *IEEE International Reliability Workshop*, pages 168–171, 2002.

[4] Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. King, and B. Nikolic. FinFET-based SRAM design. *IEEE International Symposium on Low Power Electronics and Design*, pages 2–7, 2005.

[5] B. Giraud, A. Amara, and A. Vladimirescu. A comparative study of 6T and 4T SRAM cells in double-gate CMOS with statistical variation. *IEEE International Symposium on Circuits and Systems*, pages 3022–3025, 2007.

[6] N. Lindert, Y.-K. Choi, L. Chang, E. Anderson, W. Lee, T.-J. King, J. Bokor, and C. Hu. Quasi-planar NMOS FinFETs with sub-100 nm gate lengths. *IEEE Device Research Conference*, pages 26–27, 2001.

[7] L. Chang, D.M. Fried, J. Hergenrother, J.W. Sleight, R.H. Dennard, R.K. Montoye, L. Sekaric, S.J. McNab, A.W. Topol, C.D. Adams, K.W. Guarini, and W. Haensch. Stable SRAM cell design for the 32 nm node and beyond. *IEEE Symposium on VLSI Technology*, pages 128–129, 2005.

[8] L. Chang, Y. Nakamura, R.K. Montoye, J. Sawada, A.K. Martin, K. Kinoshita, F.H. Gebara, K.B. Agarwal, D.J. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek. A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS. *IEEE Symposium on VLSI Circuits*, pages 252–253, 2007.

[9] Z. Liu and V. Kursun. High read stability and low leakage cache memory cell. *IEEE International Symposium on Circuits and Systems*, pages 2774–2777, 2007.

[10] B. Calhoun and A. Chandrakasan. A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation. *IEEE Journal of Solid-State Circuits*, pages 680–688, 2007.

[11] T.-H. Kim, J. Liu, J. Keane, and C. Kim. A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme. *IEEE International Solid-State Circuits Conference*, pages 330–332, 2007.

[12] D.M. Fried, J.M. Hergenrother, A.W. Topol, L. Chang, L. Sekaric, J.W. Sleight, S.J. McNab, J. Newbury, S.E. Steen, G. Gibson, Y. Zhang, N.C.M. Fuller, J. Bucchignano, C. Lavoie, C. Cabral Jr, D. Canaperi, O. Dokumaci, D.J. Frank, E.A. Duch, I. Babich, K. Wong, J.A. Ott, C.D. Adams, T.J. Dalton, R. Nunes, D.R. Medeiros, R. Viswanathan, M. Ketchen, M. Ieong, W. Haensch, and K.W. Guarini. Aggressively scaled (0.143 /spl mu/m/sup 2/) 6T-SRAM cell for the 32 nm node and beyond. *IEEE International Electron Devices Meeting*, pages 261 – 264, 2004.

[13] M. Okuno, K. Okabe, T. Sakuma, K. Suzuki, T. Miyashita, T. Yao, H. Morioka, M. Terahara, Y. Kojima, H. Watatani, K. Sugimoto, T. Watanabe, Y. Hayami, T. Mori, T. Kubo, Y. Iba, I. Sugiura, H. Fukutome, Y. Morisaki, H. Minakata, K. Ikeda, S. Kishii, N. Shimizo, T. Tanaka, S. Asai, M. Nakaishi, S. Fukuyama, A. Tsukune, M. Yamabe, I. Hanyuu, M. Miyajima, M. Kase, K. Watanabe, S. Satoh, and T. Sugii. 45nm node CMOS integration with a novel STI structure and full-NCS/Cu interlayers for low-operation-power (LOP) applications. *International Electron Devices Meeting*, pages 52–55, 2005.

[14] H. Nii, T. Sanuki, Y. Okayama, K. Ota, T. Iwamoto, T. Fujimaki, T. Kimura, R. Watanabe, T. Komoda, A. Eiho, K. Aikawa, H. Yamaguchi, R. Morimoto, K. Ohshima, T. Yokoyama, T. Matsumoto, K. Hachimine, Y. Sogo, S. Shino, S. Kanai, T. Yamazaki, S. Takahashi, H. Maeda, T. Iwata, K. Ohno, Y. Takegawa, A. Oishi, M. Togo, K. Fukasaku, Y. Takasu, H. Yamasaki, H. Inokuma, K. Matsuo, T. Sato, M. Nakazawa, T. Katagiri, K. Nakazawa, T. Shinyama, T. Tetsuka, S. Fujita, Y. Kagawa, K. Nagaoka, S. Muramatsu, S. Iwasa, S. Mimotogi, K. Yoshida, K. Sunouchi, M. Iwai, M. Saito, M. Ikeda, Y. Enomoto, H. Naruse, K. Imai, S. Yamada, N. Nagashima, T. Kuwata, and F. Matsuoka. A 45nm high performance bulk logic platform technology (CMOS6) using ultra high NA(1.07) immersion lithography with hybrid dual-damascene structure and porous low-k BEOL. *IEEE International Electron Devices Meeting*, pages 685–688, 2006.

[15] S.-Y. Wu, C.W. Chou, C.Y. Lin, M.C. Chiang, C.K. Yang, M.Y. Liu, L.C. Hu, C.H. Chang, P.H. Wu, C.I. Lin, H.F. Chen, S.Y. Chang, S.H. Wang, P.Y. Tong, Y.L. Hsieh, P.Y. Tong, J.J. Liaw, K.H. Pan, C.H. Hsieh, C.H. Chen, J.Y. Cheng, C.H. Yao, W.K. Wan, T.L. Lee, K.T. Huang, C.C Chen, K.C. Lin, L.Y. Yeh, K.C. Ku, S.C. Chen, C.W. Chang, H.J. Lin, S.M. Jang, Y.C. Lu, J.H. Shieh, M.H. Tsai, J.Y. Song, K.S. Chen, V. Chang, S.M. Cheng, S.H. Yang, C.H. Diaz, Y.C. See, and M.S. Liang. A 32nm CMOS low power SoC platform technology for foundry applications with functional high density SRAM. *International Electron Devices Meeting*, pages 263–266, 2007.

[16] Y. Yamagata. Embedded memory technology for low power systems. Adapted from *IEEE International Electron Devices Meeting*, 2005. Short Course Presentation.

[17] E. Seevinck, F. List, and J. Lohstroh. Static-noise margin analysis of MOS SRAM cells. *IEEE Journal of Solid-State Circuits*, pages 748–754, 1987.

[18] S.-F. Huang, C.-Y. Lin, Y.-S. Huang, T. Schafbauer, M. Eller, Y.-C. Cheng, S.-M. Cheng; S. Sportouch, W. Jin, N. Rovedo, A. Grassmann, Y. Huang, J. Brighten, C.H. Liu, B. von Ehrenwall, N. Chen, J. Chen; O.S. Park, M. Commons, A. Thomas, M.-T. Lee, S. Rauch, L. Clevenger, E. Kaltalioglu, P. Leung, J. Chen, T. Schiml, and C. Wann. High performance 50 nm cmos devices for microprocessor and embedded processor core applications. *IEEE International Electron Devices Meeting*, pages 237–240, 2001.

[19] M. Khare, S. H. Ku, R.A. Donaton, S. Greco, C. Brodsky, X. Chen, A. Chou, R. DellaGuardia, S. Deshpande, B. Doris, S.K.H. Fung, A. Gabor, M. Gribelyuk, S. Holmes, F.F. Jamin, W.L. Lai, W.H. Lee, Y. Li, P. McFarland R. Mo, S. Mittl, S. Narasimha, D. Nielsen, R. Purtell, W. Rausch, S. Sankaran, J. Snare, L. Tsou, A. Vayshenker, T. Wagner, D. Wehella-Gamage, E. Wu, S. Wu, W. Yan, E. Barth, R. Ferguson, P. Gilbert, D. Schepis, A. Sekiguchi, R. Goldblatt, J. Welser, K.P. Muller, and P. Agnello. A high performance 90nm SOI technology with 0.992 μm^2 6T-SRAM cell. *IEEE International Electron Devices Meeting*, pages 407–410, 2002.

[20] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. Mclntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, and M. Bohr. A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors. *IEEE International Electron Devices Meeting*, pages 978–980, 2003.

[21] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, I. Jeong, C. Kenyan, E. Lee, S-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigenvald, S. Tyagi, C. Weber, B. Woolely, A. Yeoh, K. Zhang, and M. Bohr. A 65nm logic technology featuring 35nm gate lengths and enhanced channel strain and 8 Cu interconnect layers and low-k ILD and 0.57 μm^2 SRAM cell. *IEEE International Electron Devices Meeting*, pages 657–660, 2004.

[22] W-H. Lee, A.Waite, H. Nii, H. M. Nayfeh, V. McGahay, H. Nakayama, D. Fried, H. Chen, L. Black, R. Bolam, J. Cheng, D. Chidambarrao, C. Christiansen, M. Cullinan-Scholl, D. R. Davies, A. Domenicucci, P. Fisher, J. Fitzsimmons, J. Gill, M. Gribelyuk, D. Harmon, J. Holt, K. Ida, M. Kiene, J. Kluth, C. Labelle, A. Madan, K. Malone, P. V. McLaughlin, M. Minami, D. Mocuta, R. Murphy, C. Muzzy, M. Newport, S. Panda, I. Peidous, A. Sakamoto, T. Sato, G. Sudo, H. VanMeer, T. Yamashita, H. Zhu, P. Agnello, G. Bronner G. Freeman, S-F Huang, T. Ivers, S. Luning, K. Miyamoto, H. Nye, J. Pellerin, K. Rim, D. Schepis, T. Spooner, X. Chen, and M. Khare. High performance 65 nm SOI technology with enhanced transistor strain and advanced-low-K BEOL. *IEEE International Electron Devices Meeting*, 2005.

[23] K. Agawa, H. Hara, T. Takayanagi, and T. Kuroda. A bit-line leakage compensation scheme for low-voltage SRAMs. *IEEE Symposium on VLSI Circuits*, pages 70–71, 2000.

[24] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano, and T. Nakano. A divided word-line structure in the static RAM and its application to a 64K full CMOS RAM. *IEEE Journal of Solid-State Circuits*, pages 479–485, 1983.

[25] M. Agostinelli, J. Hicks, J. Xu, B. Woolery, K. Mistry, K. Zhang, S. Jacobs, J. Jopling, W. Yang, B. Lee, T. Raz, M. Mehalel, P. Kolar, Y. Wang, J. Sandford, D. Pivin, C. Peterson, M. DiBattista, S. Pae, M. Jones, S. Johnson, and G. Subramanian. Erratic fluctuations of SRAM cache vmin at the 90nm process technology node. *IEEE International Electron Devices Meeting*, pages 655–658, 2005.

[26] R. Rodriguez, R. V. Joshi, J. H. Stathis, and C. T. Chuang. Oxide breakdown model and its impact on SRAM cell functionality. *International Conference on Simulation of Semiconductor Processes and Devices*, pages 283–286, 2003.

[27] H. Kobayashi, K. Shiraishi, H. Tsuchiya, M. Motoyoshi, H. Usuki, Y. Nagai, K. Takahisa, T. Yoshiie, Y. Sakurai, and T. Ishizaki. Soft errors in SRAM devices induced by high energy neutrons and thermal neutrons and alpha particles. *IEEE International Electron Devices Meeting*, pages 337–340, 2002.

[28] G. Gasiot, D. Giot, and P. Roche. Alpha-induced multiple cell upsets in standard and radiation hardened SRAMs manufactured in a 65nm CMOS technology. *IEEE Transactions on Nuclear Science*, pages 3479–3486, 2006.

[29] D. Burnett, K. Erington, C. Subramanian, and K. Baker. Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits. *IEEE Symposium on VLSI Technology*, pages 15–16, 1994.

[30] M. Pelgrom, A. Duinmaijer, and A. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, pages 1433–1440, 1989.

[31] K. Takeuchi, T. Fukai, T. Tsunomura, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto. Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies. *International Electron Devices Meeting*, pages 467–470, 2007.

[32] T. Yamaguchi, H. Namatsu, M. Nagase, K. Yamazaki, and K. Kurihara. Nanometer-scale linewidth fluctuations caused by polymer aggregates in resist films. *Applied Physics Letters*, pages 2388–2390, 1997.

[33] A. Balasinski and D. Coburn. Comparison of mask writing tools and mask simulations for 0.16 μm devices. *IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, pages 372–377, 1999.

[34] X. Ouyang, T. Deeter, C.N. Berglund, R.F.W. Pease, J. Lee, and M.A. McCord. High-throughput high-density mapping and spectrum analysis of transistor gate length variations in SRAM circuits. *IEEE Transactions on Semiconductor Manufacturing*, pages 318–329, 2001.

[35] S.-M. Jung, H. Kwon, J. Jeong, W. Cho, S. Kim, H. Lim, K. Koh, Y. Rah, J. Park, H. Kang, G. Lyu, J. Park, C. Chang, Y. Jang, D. Park, K. Kim, and M.-Y. Lee. A novel 0.79 μm^2 SRAM cell by KrF lithography and high performance 90 nm CMOS technology for ultra high speed SRAM. *IEEE International Electron Devices Meeting*, pages 419–422, 2002.

[36] S. Thompson, M. Alavi, R. Arghavani, A. Brand R. Bigwood, J. Brandenburg, B. Crew, V. Dubin, M. Hussein, P. Jacob, C. Kenyon, E. Lee, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, M. Prince, R. Schweinfurth, S. Sivakumar, P. Smith, M. Stettler, S. Tyagi, M. Wei, J. Xu, S. Yang, and M. Bohr. An enhanced 130nm generation logic technology featuring 60nm transistors optimized for high performance and low power at 0.7 - 1.4 V. *IEEE International Electron Devices Meeting*, pages 257–260, 2001.

[37] Y. Fukaura, K. Kasai, Y. Okayama, H. Kawasaki, K. Isobe, M. Kanda, K. Ishimaru, and H. Ishiuchi. A highly manufacturable high density embedded SRAM technology. *IEEE International Electron Devices Meeting*, pages 415–418, 2002.

[38] C. B. Oh, H. S. Kang, H. J. Ryu, M. H. Oh, H. S. Jung, Y. S. Kim, J. H. Lee, N. I. Lee, K. H. Cho, D. H. Lee, T. H. Yang, I. S. Cho, H. K. Kang, Y. W. Kim, and K. P. Suh. Manufacturable embedded CMOS 6T-SRAM technology with high-k gate dielectric device for system-on-chip applications. *IEEE International Electron Devices Meeting*, pages 423–426, 2002.

[39] Y. Hirano, T. Ipposhi, H. Dang, T. Matsumoto, T. Iwamatsu, K. Nii, Y. Tsukamoto, T. Yoshizawa, H. Kato, S. Maegawa, K. Arimoto, Y. Inoue, M. Inuishi, and Y. Ohji. Impact of actively body-bias controlled (ABC) SOI SRAM by using direct body contact technology for low-voltage application. *IEEE International Electron Devices Meeting*, pages 35–38, 2003.

[40] M. Ball, J. Rosal, R. McKee, WK Loh, T. Houston, R. Garcia, J. Raval, D. Li, R. Hollingsworth, R. Gury, R. Eklund, J. Vaccani, B. Castellano, F. Piacibello, S. Ashburn, A. Tsao, A. Krishnan, J. Ondrusek, and T. Anderson. A screening methodology for vmin drift in SRAM arrays with application to sub-65nm nodes. *IEEE International Electron Devices Meeting*, pages 705–708, 2006.

[41] R. Venkatraman, R. Castagnetti, and S. Ramesh. The statistics of device variations and its impact on SRAM bitcell performance and leakage and stability. *International Symposium on Quality Electronic Design*, 2006.

[42] M. Yamaoka and H. Onodera. A detailed Vth-variation analysis for sub-100-nm embedded SRAM design. *International System-on-Chip Conference*, pages 315–318, 2006.

[43] A. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin. Fluctuation limits and scaling opportunities for CMOS SRAM cells. *IEEE International Electron Devices Meeting*, pages 659–662, 2005.

[44] R. Heald and P. Wang. Variability in sub-100nm SRAM designs. *International Conference on Computer Aided Design*, pages 347–352, 2004.

[45] D. Burnett. Statistical design issues of SRAM bitcells and sense amps. *IEEE Silicon on Insulator Conference*, 2006. Short Course.

[46] T. Ichikawa and M. Sasaki. A new analytical model of SRAM cell stability in low-voltage operation. *IEEE Transactions on Electron Devices*, pages 54–61, 1996.

[47] Q. Chen, A. Guha, and K. Roy. An accurate analytical SNM modeling technique for SRAMs based on butterworth filter function. *IEEE International Conference on VLSI Design*, pages 615–620, 2007.

[48] B. Calhoun and A. Chandrakasan. Static noise margin variation for sub-threshold SRAM in 65-nm CMOS. *IEEE Journal of Solid-State Circuits*, pages 1673–1679, 2006.

[49] C. Wann, R. Wong, D. Frankt, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright, and J. Sudijono. SRAM cell design for stability methodology. *IEEE VLSI-TSA International Symposium*, pages 21–22, 2005.

[50] C.-K. Tsai and M. Marek-Sadowska. Analysis of process variation's effect on SRAM's read stability. *International Symposium on Quality Electronic Design*, 2006.

[51] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, pages 1859–1880, 2005.

[52] K. Agarwal and S. Nassif. Statistical analysis of SRAM cell stability. *International Symposium on Quality Electronic Design*, 2006.

[53] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake. Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit. *International Solid State Circuits Conference*, page 34.5, 2006.

[54] B. Calhoun and A. Chandrakasan. Analyzing static noise margin for sub-threshold SRAM in 65nm CMOS. *European Solid-State Circuits Conference*, pages 363–366, 2005.

[55] TAURUS is a trademark of Synopsys, Inc.

[56] A. Carlson, Z. Guo, S. Balasubramanian, L.-T. Pang, T.-J. King, and B. Nikolic. FinFET SRAM with enhanced read / write margins. *IEEE Silicon on Insulator Conference*, pages 105–106, 2006.

[57] H. Ananthan and K. Roy. Technology and circuit design considerations in quasi-planar double-gate SRAM. *IEEE Transactions on Electron Devices*, pages 242–250, 2006.

[58] S.-H. Kim and J. Fossum. Design optimization and performance projections of double-gate FinFETs with gatesource/drain underlap for SRAM application. *IEEE Transactions on Electron Devices*, pages 1934–1942, 2007.

[59] K. Takeuchi, R. Koh, and T. Mogami. A study of the threshold voltage variation for ultra-small bulk and SOI CMOS. *IEEE Transactions on Electron Devices*, pages 1995–2001, 2001.

[60] K. Samsudin, B. Cheng, A.R. Brown, S. Roy, and A. Asenov. UTB SOI SRAM cell stability under the influence of intrinsic parameter fluctuation. *European Solid State Device Engineering Research Conference*, pages 553–556, 2005.

[61] A. Dixit, K. G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, and K. De Meyer. Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: Role of line-edge-roughness. *IEEE International Electron Devices Meeting*, pages 709–712, 2006.

[62] Y. Okayama, T. Saito, K. Nakajima, S. Taniguchi, T. Ono, K. Nakayama, R. Watanabe, A. Oishi, A. Eiho, T. Komoda, T. Kimura, M. Hamaguchi, Y. Takegawa, T. Aoyama, T. Iinuma, K. Fukasaku, R. Morimoto, K. Oshima, K. Oono, M. Saito, M. Iwai, N. Nagashima, and F. Matsuoka. Suppression effects of threshold voltage variation with Ni FUSI gate electrode for 45nm node and beyond LSTP and SRAM devices. *IEEE Symposium on VLSI Technology*, pages 96–97, 2006.

[63] S. Mukhopadhyay, K. Kim, H. Mahmoodi, and K. Roy. Design of a process variation tolerant self-repairing SRAM for yield enhancement in nanoscaled CMOS. *IEEE Journal of Solid-State Circuits*, pages 1370–1382, 2007.

[64] M. Sumita, S. Sakiyama, M. Kinoshita, Y. Araki, Y. Ikeda, and K. Fukuoka. Mixed body bias techniques with fixed Vt and Ids generation circuits. *IEEE Journal of Solid-State Circuits*, pages 60–66, 2005.

[65] H. Morimura and N. Shibata. A step-down boosted-wordline scheme for 1-v battery-operated fast SRAM's. *IEEE Journal of Solid-State Circuits*, pages 1220–1227, 1998.

[66] H. Pilo, J. Barwin, G. Braceras, C. Browning, S. Burns, J. Gabric, S. Lamphier, M. Miller, A. Roberts, and F. Towler. An SRAM design in 65nm and 45nm technology nodes featuring read and write-assist circuits to expand operating voltage. *IEEE Symposium on VLSI Circuits*, pages 15–16, 2006.

[67] P. Elakkumanan, J. B. Kuang, K. Nowka, R. Sridhar, R. Kanj, and S. Nassif. SRAM local bit line access failure analyses. *International Symposium on Quality Electronic Design*, 2006.

Chapter 2

Understanding Variation in SRAM

2.1 Introduction

Effective reduction of variation in SRAM metrics requires a thorough understanding of its origins. Although measured SRAM variations have been linked generally to process variations, it is not initially obvious exactly how such variations cause failures. Do all variations matter equally, on all devices? Are correlated variations significant, or do mismatch variations dominate? These kinds of questions require accurate modeling of SRAM metrics down to the device parameters. Understanding the mechanisms of how parameter variation affects these metrics can inform cell and array design and improve SRAM performance and yield.

SRAM variability has become so significant of a concern that it now influences device and technology design. Novel processes have been presented to reduce variation caused by line edge roughness [1] and dopant penetration [2]. Gate length scaling in SRAM has slowed to reduce variability further. To gauge the effectiveness of design options at this level, a model is desired that can estimate SRAM metrics without the need for process development and characterization. To estimate potential yield, it must be able to simulate quickly across a wide range of perturbations. Although the mixed-mode capabilities of a device simulator could provide excellent accuracy across such a range, these simulators

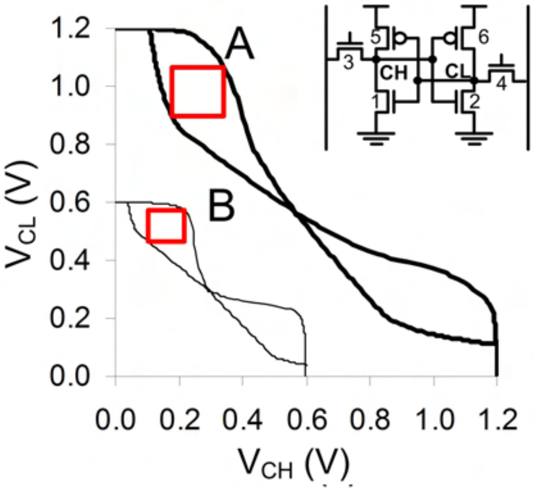

Table 2.1. Device Operating Modes

Device	A ($V_{DD} = 1.2V$)	B ($V_{DD} = 0.6V$)
PD1	linear	linear
PD2	saturation	subthreshold
PG3	saturation	weak saturation
PG4	subthreshold	subthreshold
PU5	subthreshold	subthreshold
PU6	linear	linear

Figure 2.1. As V_{DD} scales down but V_T stays constant, the operating modes of the SRAM devices change, requiring new equations to represent the voltage transfer curves. Points A and B illustrate the change at two operating points relevant to calculating SNM. In particular, the transition of PD2 into subthreshold requires subthreshold I-V modeling at low V_{DD}.

provide far more information than what is required and are notoriously slow for it. A better solution is a model that accurately represents SRAM metrics as a function of individual device parameters. For ideal speed, the model equations should be in closed form or at least require only a minimal amount of iteration.

Under constant field scaling, such a model was feasible. Seevinck *et al.* presented a model derived from the long channel I-V equations of the square law, Eqns. 1.1 [3]. With that model, SNM could be expressed as an equation of basic device widths and lengths by solving for the butterfly curves directly. The model made several approximations, which have since proved obsolete, including ones for the operating modes of the transistors. Fig. 2.1 illustrates how operating modes can change for two points relevant to SNM calculation.

The mode of operation determines which of Eqns. 1.1 is used. A change in modes requires the derivation of a new expression for SNM. Although the algebra is tedious, a closed-form expression can be achieved for many cases; however, the accuracy is significantly degraded when subthreshold current becomes significant. This is the case in Fig. 2.1 for **PD2**, which transitions into subthreshold operation. Assuming a strict cutoff of $I_{DS} = 0$ in subthreshold (*e.g.* as in [4]) distorts the shoulder of the butterfly curves around point **B**, resulting in a 19% overestimation of SNM.

To accurately estimate SNM for this case, two adjustments must be made. Subthreshold or, specifically, weak-inversion current must be modeled around V_T. This has an exponential dependence, which diminishes the number of cases with a closed-form solution. Secondly, threshold voltage must be treated separately for each device, with dependencies on individual parameter variations. A drain bias dependence must also be included for devices with significant short channel effects.

Recent models have therefore struggled to provide fast and accurate estimates for SNM. Calhoun and Chandrakasan solved SNM for deep subthreshold operation only, and for only minor parameter variations [5]. Chen *et al.* introduced a model for the butterfly curves using the Butterworth filter function to sidestep the complicated equations; however, the accuracy of the derived SNM is limited and the butterfly curves are divorced from the device parameter dependencies [6]. In spite of this, the authors rightly observe that accurate SNM modeling does not require accuracy in all sections of the butterfly curves. Only four parts of the butterfly curves are important for accurate SNM modeling: the point **A** or **B** from Fig. 2.1, its complement on the lower half of the square, and the corresponding points on the opposite lobe.

2.2 Model Development and Validation

This work proposes a semi-analytical model to provide simultaneous fast and accurate estimates for SRAM metrics, including SNM. Rather than generating several equations to approximate SNM in various limited regions or generating approximations of the butterfly curves, this model generates an analytical expression for device I-V behavior. The butterfly curves are generated through iterated, numerical solution. This approach is similar to that employed by circuit simulators such as SPICE; however, the inputs consist of only a few device I-V targets, rather than an advanced deck of hundreds of parameters. It therefore can be fit to a device technology with less characterization. From these targets, a limited number of parameters for short-channel I-V equations are calculated, such that the model is guaranteed to be accurate at every target.

Short-channel I-V equations are chosen to provide the model with a generic basis on device physics, including effects that correspond to channel length modulation, drain-induced-barrier-lowering (DIBL), velocity saturation, and bulk charge effects (adapted from [7]) .

$$I_{DS} = \begin{cases} \mu_s C_{ox} \frac{W}{2mL} \frac{(V_{GS}-V_T)^2}{1+\frac{V_{GS}-V_T}{E_{sat}L}} (1+\lambda V_{DS}) + I_{sub}\left(1-e^{\frac{V_{DS}}{V_{th}}}\right) & V_{GS} > V_T \text{ and} \\ & V_{DS} \geq \frac{V_{GS}-V_T}{m} \\ \mu_l C_{ox} \frac{W}{L} \frac{V_{DS}(V_{GS}-V_T-\frac{mV_{DS}}{V_0})}{1+\frac{V_{GS}-V_T}{E_{sat}L}} (1+\lambda V_{DS}) + I_{sub}\left(1-e^{\frac{V_{DS}}{V_{th}}}\right) & V_{GS} > V_T \text{ and} \\ & V_{DS} < \frac{V_{GS}-V_T}{m} \\ I_{sub}\left(1-e^{\frac{V_{DS}}{V_{th}}}\right) e^{\frac{V_{GS}-V_T}{S}} & V_{GS} \leq V_T \end{cases} \tag{2.1}$$

where C_{ox} is the gate oxide capacitance per unit area, W is the width of the device, L is the gate length, V_{GS} is the voltage on the gate with respect to the source, V_{DS} is the voltage on the drain with respect to the source, I_{sub} is the constant current definition for V_T, and V_T is the threshold voltage of the transistor as a function of drain bias:

$$V_T = V_{T0} - DV_{DS} \tag{2.2}$$

The other parameters are used for fitting. Separate carrier mobilities μ_l and μ_s are used for linear and saturation, respectively, to improve the fit. To ensure continuity between operating modes, a parameter V_0 is introduced such that

$$V_0 = \frac{1}{1-\frac{\mu_s}{2\mu_l}} \tag{2.3}$$

λ is a fitting parameter corresponding to channel length modulation, D represents DIBL, E_{sat} determines the amount of velocity saturation, and S represents the subthreshold swing. For ideal MOSFETs, the parameters S and m are equivalent and represent the degree to which the gate has control of the channel. In this work, a separate, global m parameter is

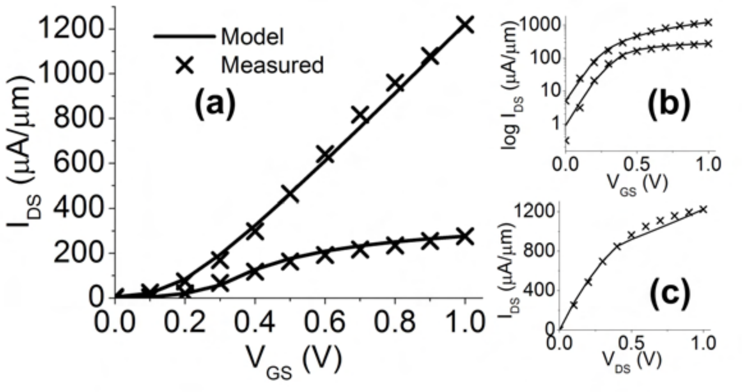

Figure 2.2. The seven parameter model introduced in this work can be used to approximate MOSFET I-V behavior even if the equations are not physically accurate. Model-generated $I_{DS} - V_{GS}$ curves (a,b) at $V_{DS} = 0.1, 1.0$V and $I_{DS} - V_{DS}$ at $V_{GS} = 1.0$V (c) exhibit good agreement with the reported I-V of a Schottky source/drain FinFET with 15nm gate length [8]. The accuracy is within 15% at all points with $I_{DS} \geq 1\mu A/\mu m$ and $0 \leq V_{DS}, V_{GS} \leq 1$V.

used to improve overall I-V agreement, and is not used to fit to individual devices. In all, there are seven independent device-specific parameters, μ_l, μ_s, λ, D, E_{sat}, S, and V_{T0}, in addition to device dimensions W and L.

To the extent that a modeled device exhibits short-channel phenomena, the I-V curves are accurate; however, the curves provide a reasonable approximation even in the presence of non-idealities or fundamental differences in carrier transport. As long as the true I-V curves of the device resemble those of a planar MOSFET, the model will be relatively accurate. This enables the model to represent advanced devices such as FinFETs without exact knowledge of the true I-V equations. Fig. 2.2 illustrates better than 15% agreement over $1\mu A/\mu m$ of this model with the reported I-V from a Schottky source/drain FinFET with 15nm gate length [8].

For the purpose of modeling SRAM, accuracy can be improved if the I-V targets around which the model is most accurate correspond to the operating biases most critical for modeling SNM and other metrics. Fig. 2.3 illustrates the biases of interest for NMOS and PMOS devices at key points on the butterfly curves. The most important regions are

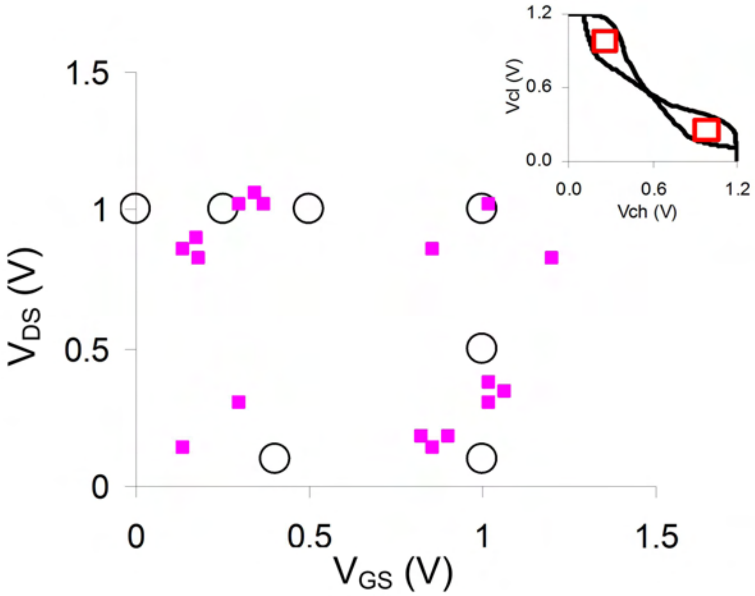

Figure 2.3. The drain and gate biases of the six SRAM devices (squares) for the key points for SNM and write-ability at $V_{DD} = 1.2V$ (inset). The important regions are at high V_{GS} or high V_{DS} and suggest locations for I-V targets (circles) to improve model accuracy. By choosing I-V targets near these key biases, the accuracy of the model is improved.

Table 2.2. Model I-V Targets

Target	I_{DLIN}	I_{DSAT}	I_{DLO}	I_{DHI}	I_{OFF}	V_{TLIN}	V_{TSAT}
V_{GS}	1.0V	1.0V	0.5V	1.0V	0.0V	N/A	N/A
V_{DS}	0.1V	1.0V	1.0V	0.5V	1.0V	0.1V	1.0V

at high V_{GS} or high V_{DS}. It is convenient that a number of commonly used I-V targets cover these regions. Table 2.2 lists the seven I-V targets used in this work.

These targets are also chosen such that there exists a one-to-one relation between them and the device parameters of Eqns. 2.1. The device parameters can then be solved as a function of the I-V targets:

$$\mu_l = \frac{\mu_{l0}}{1 - V_{TLIN} - 0.1\frac{m}{V_0}} \tag{2.4}$$

$$\text{where } \mu_{l0} = \left[I_{DLIN} - I_{sub}\left(1 - e^{-0.1/V_{th}}\right)\right]\frac{1 + \frac{1-V_{TLIN}}{E_{sat}L}}{0.1C_{ox}\left(1+0.1\lambda\right)} \tag{2.5}$$

$$\text{and } V_0 = \frac{1 - \frac{\mu_s}{2\mu_{l0}}(1 - V_{TLIN})}{1 - \frac{\mu_s}{0.2m\mu_{l0}}} \tag{2.6}$$

$$\mu_s = \frac{2m\left(I_{DSAT} - I_{sub}\right)\left(1 + \frac{1-V_{TSAT}}{E_{sat}L}\right)}{C_{ox}\left(1 - V_{TSAT}\right)^2\left(1+\lambda\right)} \tag{2.7}$$

$$\lambda = \frac{\frac{I_{DSAT}-I_{sub}}{I_{DHI}-I_{sub}} - \frac{(1-V_{TSAT})^2}{(1+0.4D-V_{TLIN})^2}\frac{E_{sat}L+1+0.4D-V_{TLIN}}{E_{sat}L+1-V_{TSAT}}}{\frac{(1-V_{TSAT})^2(E_{sat}L+1+0.4D-V_{TLIN})}{(1+0.4D-V_{TLIN})^2(E_{sat}L+1-V_{TSAT})} - \frac{I_{DSAT}-I_{sub}}{2(I_{DHI}-I_{sub})}} \tag{2.8}$$

$$D = \frac{V_{TLIN} - V_{TSAT}}{0.9} \tag{2.9}$$

$$E_{sat} = \frac{(0.5 - V_{TSAT})(I_{DLO} - I_{sub})\left(\frac{1-V_{TSAT}}{0.5-V_{TSAT}}\right)^2 - (1 - V_{TSAT})(I_{DSAT} - I_{sub})}{I_{DSAT} - I_{sub} - \left(I_{DLO} - I_{sub}\right)\left(\frac{1-V_{TSAT}}{0.5-V_{TSAT}}\right)^2} \tag{2.10}$$

$$S = -\frac{V_{TSAT}}{\ln(\frac{I_{OFF}}{I_{sub}})} \tag{2.11}$$

$$V_{T0} = V_{TLIN} + 0.1D \tag{2.12}$$

For an SRAM cell, the voltage transfer characteristics are generated by balancing the currents at the internal nodes. For example, during a read operation, current flows out of the internal node **CL** through **PD2** and into the node through **PG4** and **PU6**. For a given voltage on the complementary node V_{CH}, the voltage V_{CL} is that which satisfies

$$\begin{aligned} I_{D2}\left(V_{GS} = V_{CH}, V_{DS} = V_{CL}\right) &= I_{D4}\left(V_{GS} = V_{\overline{BL}} - V_{CL}, V_{DS} = V_{\overline{BL}} - V_{CL}\right) \\ &\quad + I_{D6}\left(V_{GS} = V_{DD} - V_{CH}, V_{DS} = V_{DD} - V_{CL}\right) \end{aligned} \tag{2.13}$$

where I_{Dx} is the drain current through device x. Eqn. 2.13 is valid for a 6-T cell in which the only current paths are between drain and source of the three devices. It can be easily modified to accommodate other cell architectures, such as 8- or 10-T designs, or other current paths, such as gate leakage. High levels of gate leakage current can degrade cell stability [9]; however, moderate levels of 100 $nA/\mu m$ at $V_{DD} = 1$V result in small (< 1 mV) changes to SNM. Even while ignoring contributions from these current paths, in many cases

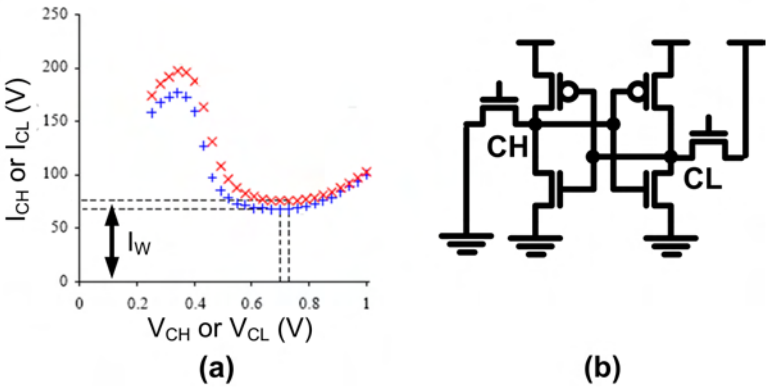

Figure 2.4. Write-ability is measured by the write-ability current, I_W, defined as the local minimum of the net current out of the internal node (a), with the cell biased as in (b). The data is extracted from measurements of a typical 90nm node SOI cell.

there is no closed-form solution to eqn. 2.13, so iteration is used. Static noise margin can be solved by rotating the voltage transfer characteristics 45 degrees and finding the local maxima, following [3].

An advantage of a numerical approach is that it can be quickly adapted to evaluate any kind of DC SRAM metric, including those for write-ability. Several write-ability metrics have been proposed, with no clear consensus on the best one [10, 11, 12, 13]. This work uses a write-ability current metric proposed by IBM, and illustrated in Fig. 2.4 [10, 14]. To determine the write-ability of one half of a cell, the cell is biased with $V_{BL} = 0$. The current I_{CH} is defined as the net current flowing out of the node, *e.g.*

$$\begin{aligned} I_{CH} &= I_{D3}\left(V_{GS}=V_{WL}, V_{DS}=V_{CH}\right) - I_{D5}\left(V_{GS}=V_{DD}-V_{CL}, V_{DS}=V_{DD}-V_{CH}\right) \\ &\quad + I_{D1}\left(V_{GS}=V_{CL}, V_{DS}=V_{CH}\right) \end{aligned} \tag{2.14}$$

Since V_{CL} is usually small in the interesting case, the last term is often neglected. I_{CH} can then be thought of as the current difference between the pass-gate device **PG3** discharging the node and the pull-up device **PU5** resisting the write. When plotted against V_{CH}, the curve takes a characteristic “N” shape. The minimum of this curve above the trip point of the inverter is the write-ability current, I_W. A higher I_W corresponds to a more write-able

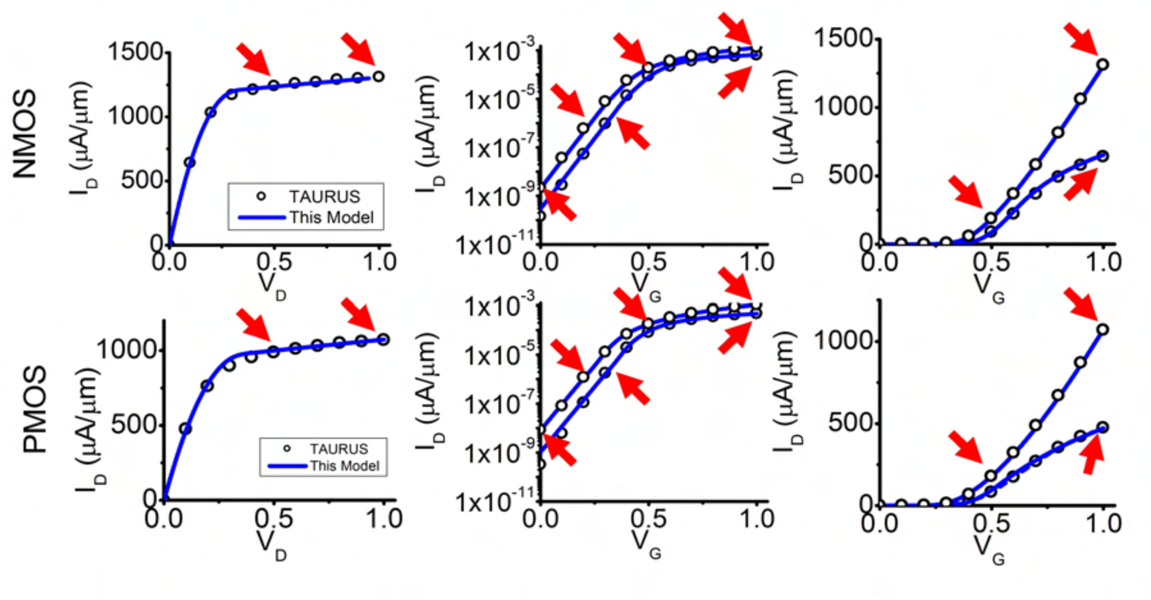

Figure 2.5. Using only seven I-V targets from TAURUS simulations of a FinFET with 22 nm gate length, the fast model presented in this work can accurately represent the full I-V behavior. The seven targets indicated by arrows are presented in Table 2.2.

cell, and $I_W \leq 0$ indicates a cell that cannot be written. As with SNM, there is an I_W corresponding to each half of the cell.

This model can therefore estimate SNM and I_W in a fast and accurate manner, even if the underlying I-V equations are not exactly correct. TAURUS device simulations are used to generate I-V curves for a FinFET with 22 nm gate length. Using these curves only at the targets of Table 2.2, the full I-V behavior of the FinFETs can be accurately modeled, for both NMOS and PMOS (Fig. 2.5). The modeled I-V relationships can also be used to solve for the voltage transfer characteristics of the SRAM cell or to generate "N" curves for write-ability analysis. The butterfly curves and write-ability curves generated from the model exhibit excellent agreement with those generated using the mixed-mode simulation capability of TAURUS (Fig. 2.6).

This model has also been validated against measurements of several hundred bitcells over a range of process options, applied biases, and temperatures. SRAM bitcells were fabricated in a "padded-out" layout, such that the internal nodes of the cells were accessible

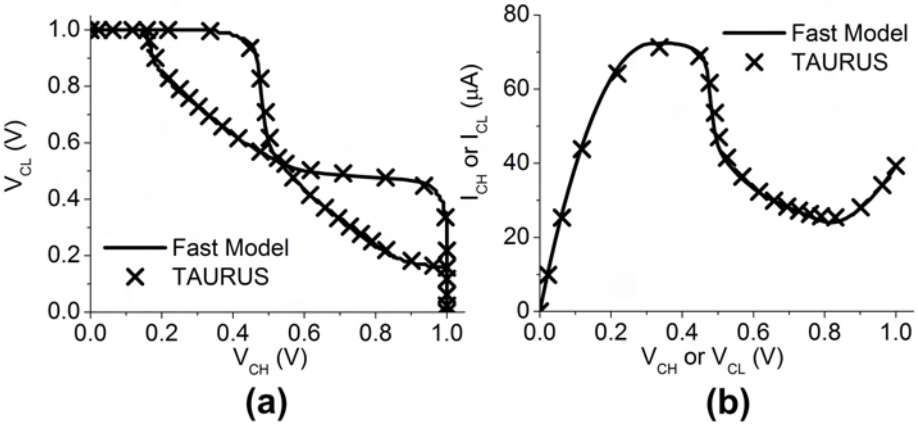

Figure 2.6. Using I-V parameters from TAURUS simulations of a FinFET with 22 nm gate length, the fast model presented in this work can predict the butterfly (a) and write-ability (b) curves for an SRAM cell. The fast model curves show excellent agreement to curves generated with the mixed-mode capability of TAURUS.

for probing. To validate the model, I-V curves were measured for each of the six transistors in the cell. SNM and I_W were then measured over a range of different biases from 0.6V to 1.2V and compared to model predictions based only on a small number of measured I-V targets: I_{DLIN}, V_{TLIN}, and V_{TSAT}. Figure 2.7 illustrates very good agreement between the model prediction and the actual measurement for a 90nm industrial process. Predictions from older SNM models are also generated to emphasize the need for this new approach [3, 4]. For the older models, a parameter representing $\mu C_{ox} W/L$ was extracted from I_{DLIN} and V_{TLIN} using eqn. 1.1. This parameter and V_{TLIN} were entered for each transistor into the models of [3, 4]. The resulting SNM predictions show very poor agreement with the measured values because short channel effects are neglected. The model still shows good agreement if the process is changed. Figure 2.8 illustrates SNM and I_W agreement under two process alterations, including the incorporation of a tensile strain layer. This adjustment strengthens the NMOS transistors relative to the pull-up devices, resulting in an improvement to I_W and a slight decrease in SNM. This model also shows promising validation results to prototype cells fabricated in a 65nm technology, even if an alternative read stability metric [10] is used (Fig. 2.9).

Overall, the accuracy of the model is very good, especially compared to other simulations

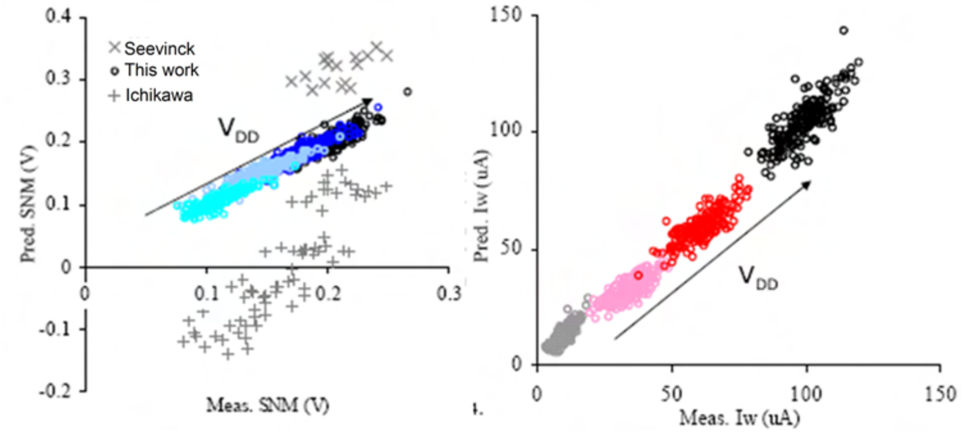

Figure 2.7. The model shows very good agreement to measured SNM and I_W from SRAM cells fabricated in a 90nm industrial process. The one-to-one agreement is much better than that of the Seevinck [3] or Ichikawa [4] models, which became obsolete with scaling as short channel effects and subthreshold currents increased.

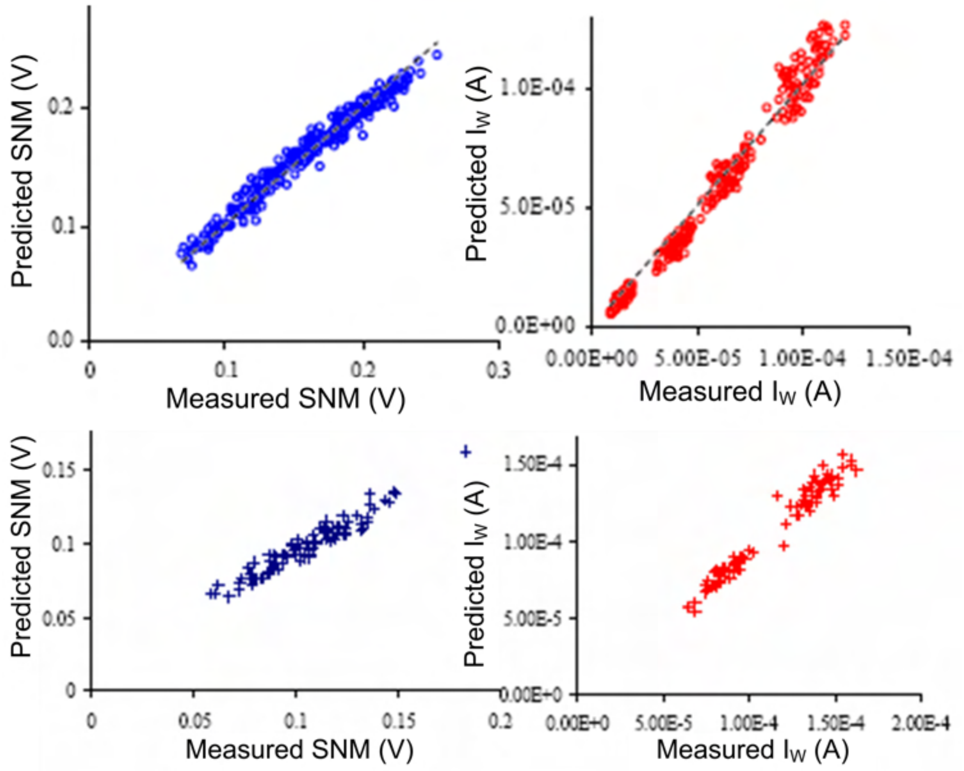

Figure 2.8. Agreement is still very good even under two separate adjustments to the fabrication process. The addition of a strain layer to enhance N-channel mobility (below) has the effect of decreasing SNM but increasing I_W.

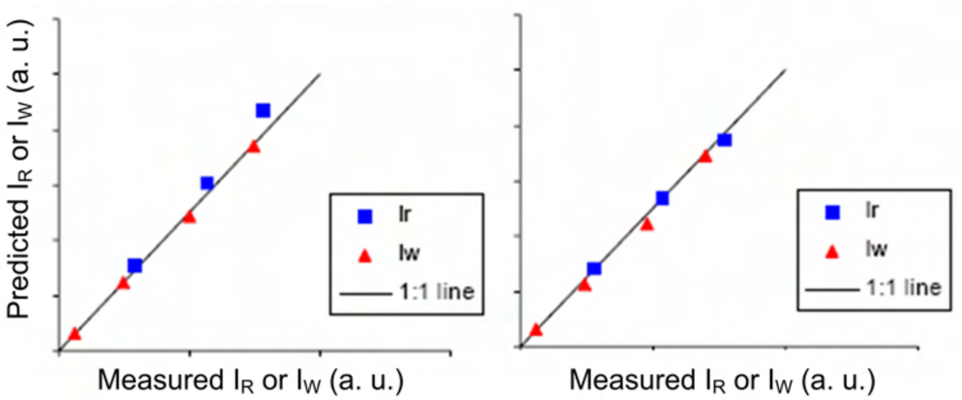

Figure 2.9. The model also shows very good agreement to measurements from an early prototype of a 65nm SOI cell. Here a read current metric (I_R) was used as an alternative to SNM. Good validation is also seen to SRAM cells fabricated in a 45nm bulk process, further validating this approach among different device technologies.

at the same speed. The accuracy is not perfect, though, and there are some scenarios under which the fit is poor. The model should not be used for cells with voltage supplies significantly higher than 1.0V, for example, since the I-V targets are fit only within the range of $0 \leq V_{DD} \leq 1.0V$. Equations 2.4 – 2.12 make assumptions about the operating modes of the I-V targets, in particular, $V_{TSAT} > 0$V and $V_{TLIN} < 0.5$V. For devices with very low or very high V_T, new I-V targets should be used or the current definition for V_T should be changed in order to meet this criterion. In practice, fitting the model to a specific device technology can require these or other small adjustments, but once fit it will provide accurate I-Vs even for devices with several standard deviations of parameter variations.

2.3 Impact of SRAM scaling trends

Besides accuracy, an advantage of this model over closed-form equations is that it can be rapidly adapted to variations in device parameters such as W, L, or V_{T0}. Small variations can be made to a parameter individually to evaluate the sensitivities of SNM and I_W. These sensitivities identify which parameters, in which devices, are most important for determining cell yield. They can illustrate the impact of device scaling or the introduction of new technologies, and they are invaluable for estimating yield.

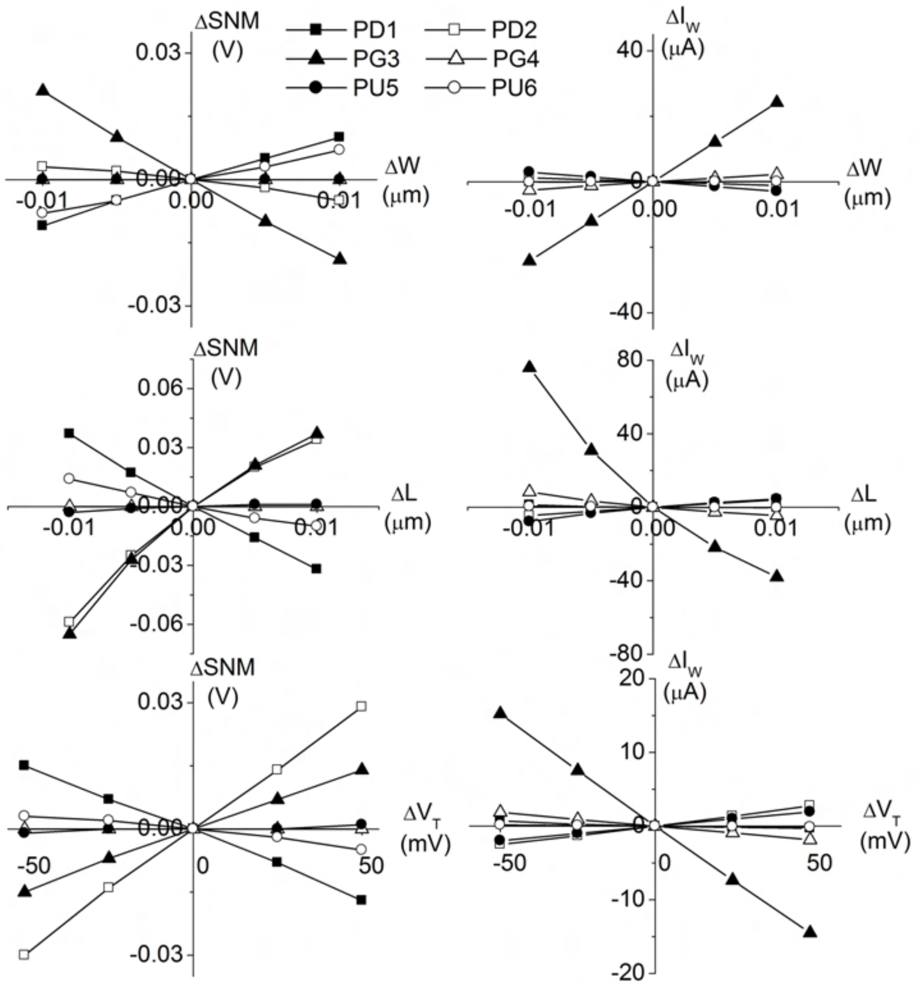

Figure 2.10. Sensitivities of SNM (left) and I_W (right) for the **CH** node, to device parameters W, L, and V_{T0} at $V_{DD} = 1.2$V.

The sensitivities of SNM and I_W to a device parameter x on device i is defined as $\partial SNM/\partial x_i$ and $\partial I_W/\partial x_i$, respectively. They can be determined from simulations by varying x_i in small amounts, as illustrated in Fig. 2.10 for the **CH** side of the cell at $V_{DD} = 1.2V$. In these simulations, a cell with I-V targets representative of a 65nm node planar SOI transistor are used (Table 2.3).

In most cases, SNM and I_W exhibit a linear response to small variations in x_i. The sensitivities are the slopes of these lines, with higher slope indicating a greater sensitivity. SNM is most sensitive to L variations in the **PD1**, **PD2**, and **PG3** devices and to V_{T0}

Table 2.3. 65nm I-V Targets and Parameter Variations

(V_T is extracted at 300 nA/μm for NMOS (PD & PG) and 70 nA/μm for PMOS (PU))

Nominal I-V Targets

	I_{DLIN} $(\mu A/\mu m)$	I_{DSAT} $(\mu A/\mu m)$	I_{DLO} $(\mu A/\mu m)$	I_{DHI} $(\mu A/\mu m)$	I_{OFF} $(nA/\mu m)$	V_{TLIN} (V)	V_{TSAT} (V)
V_{DS}	0.1V	1.0V	1.0V	0.5V	1.0V	0.1V	1.0V
V_{GS}	1.0V	1.0V	0.5V	1.0V	0.0V		
PD	700	1600	220	1400	30	0.30	0.21
PG	700	1600	220	1400	30	0.30	0.21
PU	100	500	90	350	30	0.25	0.16

Parameter Dimensions and Standard Deviations

	W (μm)	L (μm)	σ_W (nm)	σ_L (nm)	σ_{VT} (mV)
PD	0.12	0.038	2.0	2.6	18
PG	0.06	0.038	2.0	3.8	25
PU	0.06	0.038	2.0	3.8	25

variations in the **PD2** device. It is also sensitive to width variations in **PG3**, to L variations in **PU6**, and to V_{T0} variations in **PD1** and **PG3**, but to a lesser extent. For the **CL** side of the cell, the sensitivities are switched between paired devices, *e.g.* $\partial SNM_{CH}/\partial L_1 = \partial SNM_{CL}/\partial L_2$.

The high sensitivities to L reflect the importance of DIBL at high V_{DD}. Variations in L_1 and L_3 change the effective V_T of these devices and thereby affect the cell's stability during a bitline discharge. An increase in L_1 weakens **PD1**, shifting the lower shoulder of the butterfly curves up and decreasing SNM (Fig. 2.11). A decrease in L_3 strengthens **PG3**, linearizing the lower shoulder of the butterfly curves and decreasing SNM. SNM is also affected by the position and shape of the upper shoulders of the butterfly curves. A smaller L_2 decreases the gain of the **PD2**-**PU6** inverter, resulting in a rounder shoulder and less SNM. A decrease in V_{T2} pulls in the upper shoulder, decreasing SNM as well.

The importance of **PD2** to SNM on the **CH** node underlies the motivation for variation-aware SRAM design. In the introduction to beta ratios in section 1.1.1, only the devices

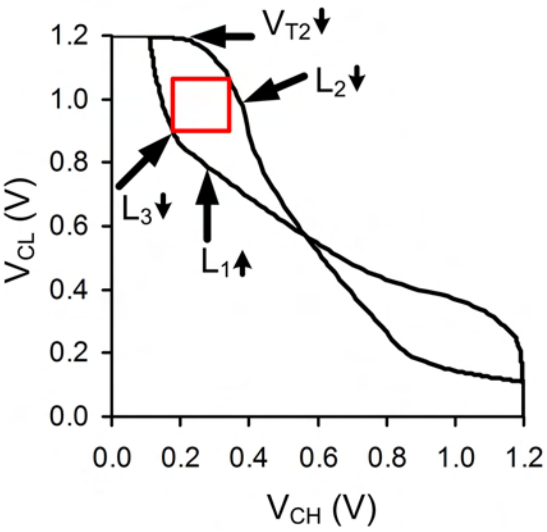

Figure 2.11. Variation in certain device parameters such as L_1, L_2, L_3, and V_{T3} can degrade SNM by changing the shape of the butterfly curves.

in the same half of the cell were considered important. Sensitivity analysis reveals that not only are the variations in both pull-down devices of comparable importance, but they have opposite effects. An increase in the strength of **PD1** is approximately as good for SNM_{CH} as it is bad for SNM_{CL}. In other words, a process variation that strengthens both the pull-down devices has less of an effect than one that causes a mismatch between them. Fig. 2.12 illustrates the SNM sensitivities to mismatch and common mode V_T variations by device type. Mismatch mode variations dominate for the **PD** devices, whereas common mode variations are negligible. This reflects the nearly opposite sensitivities of **PD1** and **PD2** in Fig. 2.10. Mismatch mode variations are of comparable importance to common mode variations in the **PG** devices, since the sensitivity to **PG4** is nearly zero. The same can be said for the **PU** devices, and the trends are similar among W and L variations as well. In the aggregate, the sensitivities to mismatch are approximately three times greater than those to common mode variations. For write-ability, the sensitivities are more evenly split between mismatch and common mode.

To some extent, this allows for the adjustment of cell SNM with a process or design change that affects all of one type of devices, such as changing an implant condition to increase NMOS V_T. The increase to SNM from **PD2** will be mostly offset by the decrease

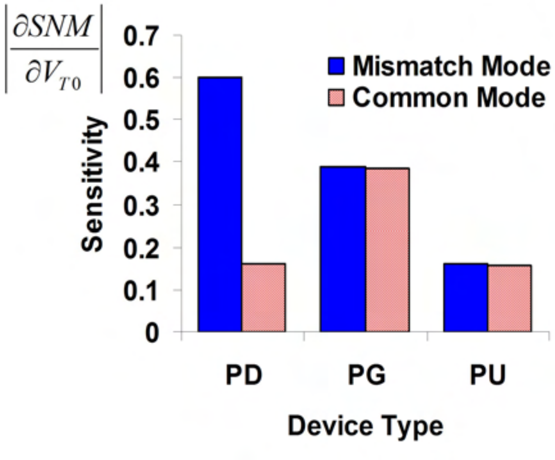

Figure 2.12. SNM exhibits much greater magnitudes of sensitivity to mismatch mode variations in the **PD** devices than to common mode variations. In the aggregate, the overall SNM sensitivity to parameter (W, L_G, and V_{T0}) mismatches are almost three times higher than those to common mode variations. For write-ability, the sensitivities are more comparable between the two modes.

from **PD1**, but the effect from **PG3** will be unopposed. Unlike with the pull-down devices, the tradeoff with pass-gate device parameters is not with the other half of the cell. The sensitivities of SNM_{CH} to **PG4** parameters is zero, so the net result will be an increase to SNM. Instead, the pass-gate tradeoff is against write-ability. Increasing pass-gate V_{T0} will decrease I_W, since the pass-gate weakens relative to the pull-up device. At high voltage, variations in pass-gate L, W, and V_{T0} dominate, and the other devices are of negligible importance. Considering the I_W definition as a difference between pass-gate and pull-up currents, one might expect the sensitivities to be shared between these devices; however, at the point where I_W is measured, at the local minimum of the I_{CH} N-curve, the inverter is switching and the gate overdrive on the pull-up is decreasing. The sensitivity of I_W to **PU5** can increase as I_W gets closer to zero, with increasing amounts of variation or lower V_{DD}, but under most conditions **PG3** will remain dominant.

The most significant effect of lowering V_{DD} is a shift in the sensitivities away from device parameters like W and L toward V_{T0} (Fig. 2.13). In part this reflects a decrease in DIBL (due to smaller drain biases) and an increase in the importance of V_T with small

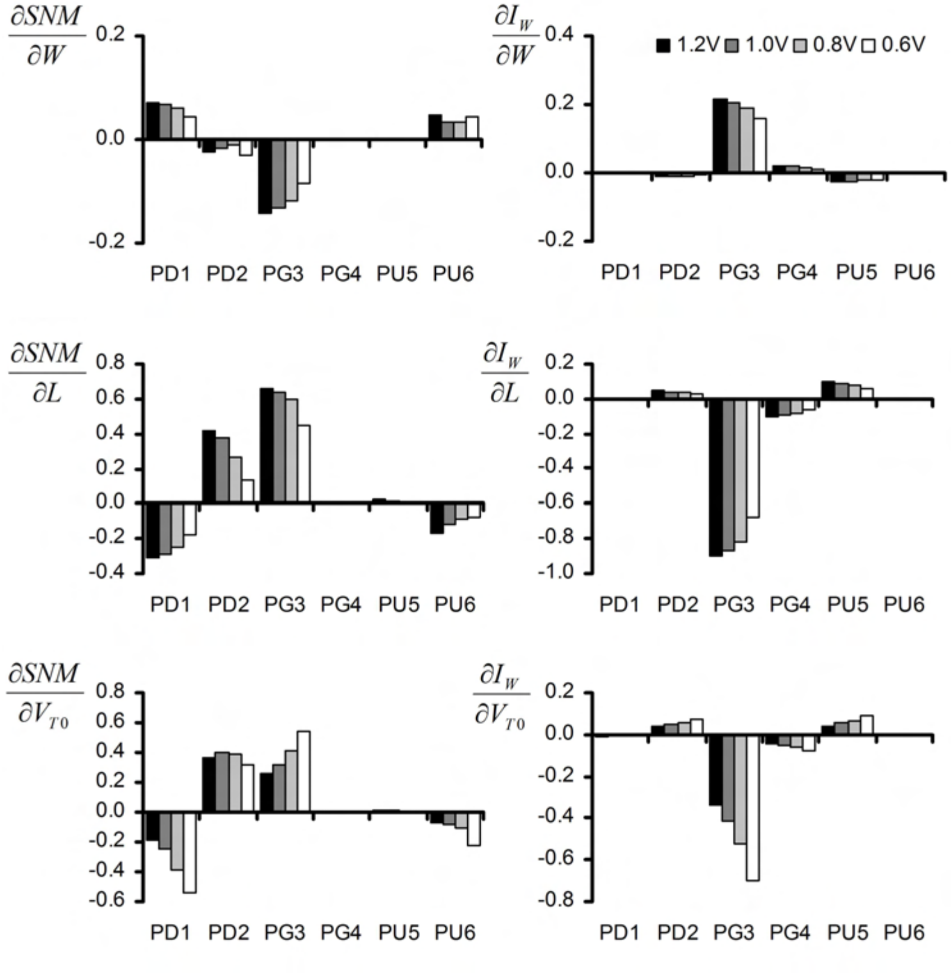

Figure 2.13. Normalized sensitivities of SNM (left) and I_W (right) for the **CH** node are a function of V_{DD}. In general, the sensitivities to W and L decrease as the effect of DIBL decreases and subthreshold currents become significant.

V_{GS}. Given saturation currents of the form $I_{DS} \propto \frac{W}{L}\left(V_{GS} - V_T\right)^2$, then

$$\frac{\partial I_{DS}/\partial V_T}{\partial I_{DS}/\partial L} = \frac{2L}{V_{GS} - V_T} \tag{2.15}$$

which gets larger as V_{GS} gets smaller. Similar effects are observed for other operating modes. It is most significant in subthreshold, where I_{DS} has an exponential dependence on V_T. To allow for a meaningful comparison among SNM, I_W, and different parameters and biases, the sensitivities in Fig. 2.13 are normalized to the standard deviations of the metric, per standard deviation of the device parameter. In other words, from the raw simulation data illustrated in Fig. 2.10,

$$\left(\frac{\partial SNM}{\partial x_i}\right)_{normalized} = \left(\frac{\partial SNM}{\partial x_i}\right)_{raw} \frac{\sigma_{xi}}{\sigma_{SNM}} \tag{2.16}$$

where σ_{xi} is the standard deviation of device parameter x on device i and

$$\sigma_{SNM}^2 = \sum_i \left(\frac{\partial SNM}{\partial x_i}\right)_{raw}^2 \sigma_{xi}^2 \tag{2.17}$$

Assuming the x_i follow independent, Gaussian distributions and that the $\frac{\partial SNM}{\partial x_i}$ are linear over the range of possible x_i, σ_{SNM}^2 is the variance of SNM for the cell. The normalized sensitivities thus have the property that

$$\sum_i \left(\frac{\partial SNM}{\partial x_i}\right)^2 = 1 \tag{2.18}$$

A similar relationship holds for I_W or any other SRAM metric.

In addition to shifting the sensitivities toward V_{T0}, a decrease in V_{DD} has several significant effects. Gate and drain leakage currents are reduced exponentially, decreasing standby power. This is the chief reason for V_{DD} scaling. The voltage transfer characteristics change in two ways: the upper shoulder of the curves becomes steeper, reflecting the high gain of the inverter in subthreshold. The lower shoulder of the curves tend to become more linear, as the inverter current decreases relative to that through the resistive load of the pass-gate device. SNM decreases, in part due to this linearization, but primarily because of the scale down in V_{DD}. Write-ability and read currents invariably decrease, and though the DC metrics can remain positive down to very low V_{DD}, in practice the cell will become too slow to access at a reasonable frequency. A common array level metric is the minimum operating voltage (V_{min}) of the array at a particular frequency. V_{min} can be constrained by any of SNM, write-ability, or read current, though with continued scaling write-ability limitations are likely to dominate.

For this reason, dual supplies for SRAM have been considered to enable continued V_{DD} scaling without degrading write-ability. A distinct wordline bias $V_{WL} > V_{DD}$ can be applied during a cell write operation, to increase the pass-gate drive current. This has a similar effect to reducing V_T dynamically, since V_T is subtracted from V_{GS} in eqns. 2.1. The cell therefore exhibits very similar sensitivities to variations or noise in V_{WL} as it does to V_{T0}. Alternatively, a second supply can be used to back bias certain transistors during the write. For example, the n-well bias can be raised during a write operation and lowered during the read [15]. One challenge for implementing a dual supply architecture is the generation of a noise-free V_{WL}.

Along with decreasing V_{DD}, the drive to scale is expected to introduce new materials and device architectures to SRAM transistors. In section 1.2, the impact of strain was briefly discussed as a complicating factor for maintaining high write-ability with minimum-width devices. The introduction of high-k dielectrics is expected to be good for SRAM, because it will reduce the standby power and the instability caused by gate leakage currents. Although achieving low V_T devices with high-k has proved challenging from a processing perspective, read stability and yield are enhanced with higher V_T devices, as long as the variation to V_T does not increase. In fact, V_T variation may decrease with a reduced equivalent oxide thickness, or if a metal gate with a specific work function is used to set V_T. This may allow for a smaller channel implant, reducing random dopant fluctuation. Random dopant fluctuation can be further decreased by a switch to fully-depleted SOI, multi-gate, or another device architecture which allows for undoped channels. In these devices, V_{T0} is set by the thickness of the channel material, which is a smaller source of variation than dopant implantation and diffusion. Such process changes may not lower the sensitivities to device parameters as much as they decrease their variances, but yield is improved in either case.

2.4 Yield Modeling and Statistical Design Methods

This modeling approach can be extended beyond nominal predictions and sensitivity analyses to projections of cell yield. There are two aspects to SRAM yield projections: how much cell variation is expected, due to known device parameter variations and the sensitivities to them, and how much cell variation a design can tolerate before failing in reading, writing, or both. These aspects are quantified in the *cell sigma*, a metric defined as the least amount of variation that causes a failure. A cell sigma can be defined for SNM, I_W, or any other metric, or for either half of a cell. If the metric follows a Gaussian distribution, the cell sigma is simply μ/σ, the mean over the standard deviation. For a metric f (*e.g.* SNM or I_W) that is not necessarily a perfect Gaussian but is subject to small, independent parameter variations x_i over a range such that f can be approximated as a linear function of x_i, the central limit theorem states that the distribution of f can be approximated as Gaussian. If so, the cell sigma is given by:

$$\text{cell sigma} = \frac{f(0)}{\sqrt{\sum_i \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_{xi}^2}} \tag{2.19}$$

Unfortunately, this simple estimation of yield is often insufficient for SRAM, where the sensitivities can become non-linear beyond several σ_{xi} of variation. This is especially true for very small dimensions, where quantum confinement or tunneling phenomena make a fundamental change to the I-V behavior of a device. Calhoun *et al.* used Monte Carlo simulations to show how the tails of a SNM distribution may deviate from Gaussian behavior, even assuming perfect Gaussian distributions for the device parameters [5]. Measurements of several hundred fabricated cells in a 45nm process also show some non-Gaussian behavior at the tails, even though the rest of the distribution looks Gaussian (Fig. 2.14). Non-Gaussian behavior at the worst-case tail of SNM has also been observed [16].

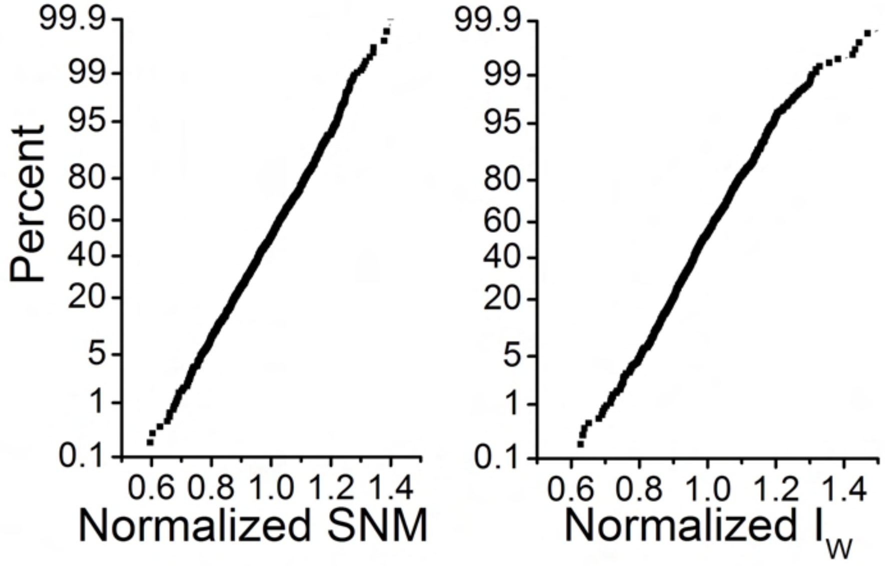

Figure 2.14. Measurements of SNM (a) and I_W (b) from 1080 SRAM cells show mostly linear behavior on a cumulative distribution plot, indicating a Gaussian distribution; however, the tails of the distribution are thought to be non-Gaussian in some cases. Data is normalized to mean.

There is no closed-form equation to accurately model yield in the face of changing sensitivities, but as before a fast and accurate estimate can be obtained with iteration. It is convenient to think of such an algorithm in a multi-dimensional variation space, wherein each dimension represents variation in a unique device and parameter combination, x_i. Increasing $|x_i|$ corresponds to increasing variation and decreasing probability of occurrence. The origin ($\vec{x} = 0$) represents the nominal design point. In this space there is a *surface of failure* for the metric f, such that every point on this surface represents a combination of device parameter variations where $f(\vec{x}) = 0$. The cell sigma is measured as the shortest distance to the surface of failure from the origin. Formally,

$$\text{cell sigma} = \min_{f(\vec{x})=0} ||\vec{x}|| \tag{2.20}$$

This is illustrated in Fig. 2.15 for the case of SNM distributed as a nearly perfect Gaussian. Although there are 18 possible dimensions of variation (W, L, and V_{T0} for six transistors), only **PD1** V_{T0} and **PG3** V_{T0} are shown. The line represents the surface of

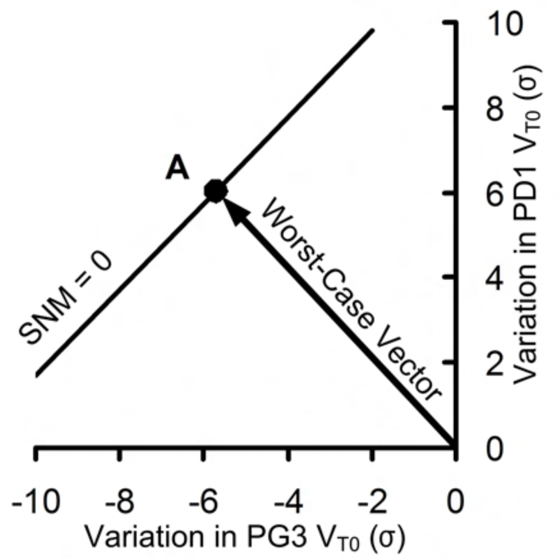

Figure 2.15. Cell sigma can be illustrated in a multi-dimensional variation space, with each dimension representing variation in a device parameter. Here only two dimensions are shown, corresponding to **PD1** V_{T0} and **PG3** V_{T0}. There is a surface to failure (SNM = 0 line) upon which the combination of parameter variations causes cell failure. The worst case vector (**A**) is the most probable (shortest distance to origin) combination of variations on this surface. The *cell sigma* is measured as the length of this vector.

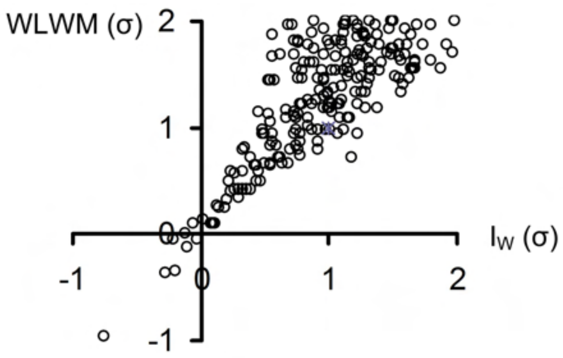

Figure 2.16. Two write-ability metrics, I_W and WLWM, have different μ/σ at high V_{DD} and only moderate correlation, as measured from cells in a 45nm process. The agreement improves dramatically at the point of zero write-ability. Yield estimations derived from simulations at the surface to failure will therefore be metric independent.

failure for SNM at $V_{DD} = 0.6$V. Although huge variations ($> 10\sigma$) in either V_{T1} or V_{T3} alone could cause SNM = 0, the most probable point is at **A**, sometimes called the worst case vector, where the line is closest to the origin. The cell sigma is the distance from that point to the origin.

An advantage of using the surface of failure to calculate cell sigma is that it is fairly metric independent. Fig. 2.16 illustrates the correlation between two write-ability metrics, I_W and a wordline-based write metric (WLWM), for 144 cells in an early industrial 45nm process. The cells are measured over a range of V_{DD} and well bias conditions. For the purposes of comparison, the points in Fig. 2.16 are normalized to the standard deviation of the 144 cells calculated at each bias condition. Although the metrics diverge at high V_{DD}–reflecting the inadequacy of μ/σ estimation–they come into agreement near zero write-ability. Simulations confirm similar behavior for other write-ability metrics. By simulating near the surface of failure, a yield estimation algorithm can be developed that is metric independent.

Fig. 2.17 illustrates an algorithm to find the worst case vector in the presence of non-linear, monotonic sensitivities using iteration. First the sensitivities are calculated around an arbitrary point $\vec{x}_n$ in variation space. As a vector, the sensitivities represent the gradient of f, $\vec{\nabla} f$, which points generally toward the surface of failure. A new point $\vec{x}'_n$ is found

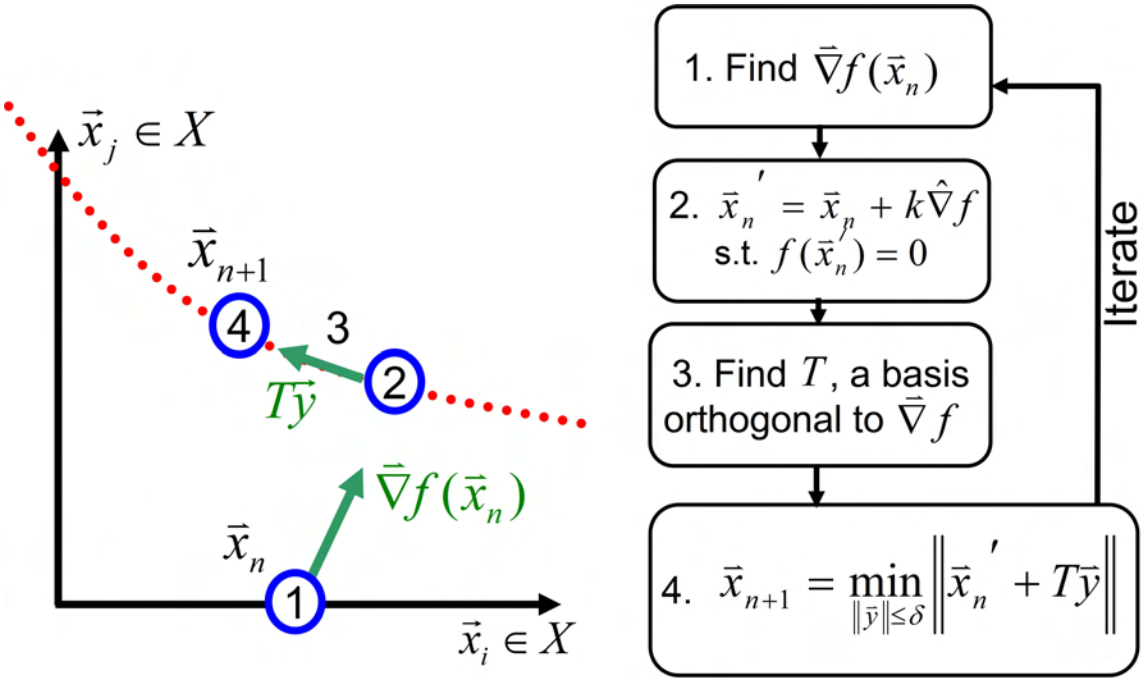
$\vec{x}_j \in X$
$\vec{x}_{n+1}$
4
3
2
1
$T\vec{y}$
$\vec{\nabla} f(\vec{x}_n)$
$\vec{x}_n$
$\vec{x}_i \in X$
1. Find $\vec{\nabla} f(\vec{x}_n)$
2. $\vec{x}_n{}' = \vec{x}_n + k\hat{\vec{\nabla}} f$ s.t. $f(\vec{x}_n{}') = 0$
3. Find T, a basis orthogonal to $\vec{\nabla} f$
4. $\vec{x}_{n+1} = \min_{\|\vec{y}\| \le \delta} \left\| \vec{x}_n{}' + T\vec{y} \right\|$
Iterate

Figure 2.17. An iterative algorithm will converge quickly to the most probable point of failure (from which the cell sigma can be determined) by using metric sensitivities.

by following this gradient out to the surface of failure. Meanwhile, a basis of vectors, T, orthogonal to $\vec{\nabla} f$ is generated. Any vector or linear combination of vectors in this basis represents a direction in variation space for which f is constant. From the point $\vec{x}'_n$ on the surface to failure, T represents all the directions that lie along that surface. A new point $\vec{x}_{n+1}$ is found in the direction from T that minimizes the distance to the origin, $||\vec{x}_{n+1}||$. Each change to a new point $\vec{x}$ is confined to a maximum distance from the previous point, over which the initial sensitivities calculations are judged to be accurate. The algorithm then repeats. It will eventually converge to the true cell sigma provided that the surface of failure is convex.

Whether or not the surface of failure is truly convex remains to be proven. A surface which is not convex could cause this algorithm to converge to a local–and not the global–minimum. Due to the number of dimensions of parameter variation, it is computationally prohibitive to grid the space and determine the convexity of the surface of failure. Since the surface of failure is not expressible by a closed-form equation, it cannot be proven convex or

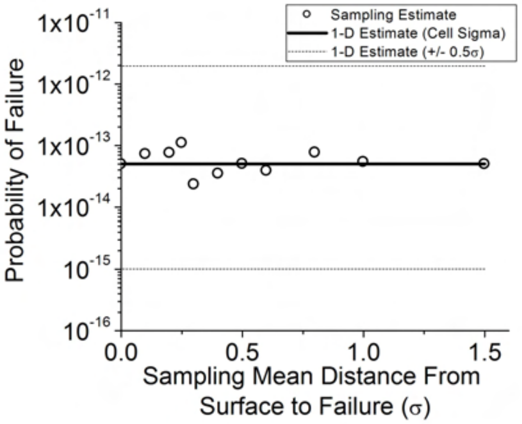

Figure 2.18. Importance sampling can be used to estimate write yield for 45nm bulk SRAM. The sampling distribution in $N = 18$ dimensions is Gaussian with N variance and different means. The yield estimate approximates the probability of a 1-D Gaussian at the cell sigma, suggesting that the surface of failure is mostly flat.

otherwise in general. In the specific case of halo-doped devices with reverse short-channel effects, the device parameters exhibit an inherent non-convexity, since V_T increases and then decreases with decreasing gate length. If the surface of failure traverses the non-convex region of this parameter, then it may be non-convex. In many cases, though, the surface of failure does appear to be convex, and convergence failures are often artifacts of the simulation. Simulations using random initial guesses often converge to the same point, suggesting at least that any non-convex portions of the surface of failure are small. The yield can be estimated from importance sampling a small distance ($\leq 1.5\sigma$) away from the surface of failure (Fig. 2.18). A decreasing estimate with sampling closer to the surface of failure would suggest a non-convex shape; however, for simulations validated to a 45nm bulk process, the estimate is consistent with a flat (1-D) surface. Therefore, it is reasonable to assume convexity for most analyses, but it is not yet proven.

If the surface of failure is convex, this approach can converge to an accurate solution much faster than Monte Carlo-based methods, which can require billions of repetitions. The key to its speed is the use of sensitivities to guide the search. Convergence can be achieved with a binary pass/fail metric, but it requires many more simulations, of the order of $O\left(n \log 1/R\right)$ versus $O\left(n - \frac{\log 1/R}{\log 1/\epsilon}\right)$ with this method, where n is the number of device

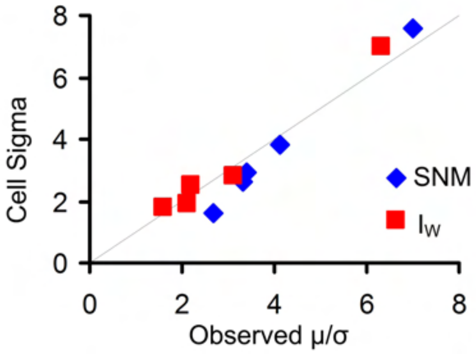

Figure 2.19. Cell sigma simulations correlate with measurements of mean to standard deviation from 144 cells fabricated in a 45nm process.

parameters, R is the resolution of the search, and ϵ $(0 < \epsilon < 1)$ is the percent error in the σ_{SNM} estimate of eqn 2.17.

By representing the most probable point of failure, the cell sigma only approximates yield. The probability of occurrence for that precise combination of variations is infinitesimally small; however, designs with higher cell sigmas can be expected generally to have higher yields. Fig. 2.19 illustrates the cell sigma simulated from I-V targets for a 45nm planar bulk design, at various V_{DD} and well biases. A positive correlation is observed with the ratio of mean to standard deviation, calculated from a measurement of 144 cells. Linear fits can be made to both read stability and write-ability; however, the slopes are different. Because cell sigma and μ/σ are only estimates of yield, only an approximate agreement is expected, which is consistent with the observations. The agreement provides further evidence that, in combination with theory, suggests a fast and accurate yield simulation can indeed be obtained with an iterative algorithm.

To use an iterative algorithm though, it is necessary to have accurate model I-Vs for devices which exhibit large amounts of variation. The safest approach is to provide I-V targets for devices with representative amounts of variation in W, L, or V_{T0}. The appropriate I-V targets for an arbitrary amount of variation can then be interpolated and the appropriate parameters extracted. Interpolating the targets may require quadratic or exponential fitting equations. This approach is best suited for device technologies which are well-characterized or well-modeled with a device simulator.

Alternatively, if the physics of the device are well-modeled with eqns. 2.1, large variations can be handled by modifying the extracted parameters directly. In general, the parameters of eqns. 2.4 – 2.12 are insensitive to W, L, or V_{T0} variations, with the exception of DIBL, which follows an exponential dependence:

$$D \approx V_{DS} e^{-L/\ell} \tag{2.21}$$

where ℓ is a constant for a given technology, dependent on the gate control of the channel. Additionally, there may be minor sensitivities of μ_l and μ_s to V_{T0}, if the bias conditions are such that dopant concentrations constrain mobilities, and S to L in extreme cases. If the parameter sensitivities can be neglected or modeled with closed-form equations, then this approach will provide faster simulations. It also avoids the small non-monotonic errors that rarely occur when re-extracting device parameters.

The final caveat to the fast modeling approach of this work is the assumption that device parameters W, L, and V_{T0} undergo independent variations. In many device architectures this is true and can be verified experimentally by correlating I-V targets across devices. It is not guaranteed for all architectures though, and some devices may exhibit weak correlations between V_{T0} and W or L if, for example, the device has a strong narrow-width effect or a steep V_T rolloff curve, respectively. In such cases it is best to revise eqn. 2.2 to explicitly account for these effects, *e.g.*

$$V_T = V_{T0}\left(1 - k_1 e^{-W/k_2}\right)\left(1 - k_3 e^{-L/k_4}\right) - DV_{DS} \tag{2.22}$$

where the k_i are constants fit to the device technology. Other equations may be used instead of 2.22 as long as each device parameter corresponds to a unique, independent random process.

As long as this condition is satisfied, read and write yield projections can be made in a matter of minutes. Fig. 2.20 illustrates read and write cell sigmas for the 65nm node cell. In general, read and write yields decrease with very low V_{DD} as both nominal SNM and I_W

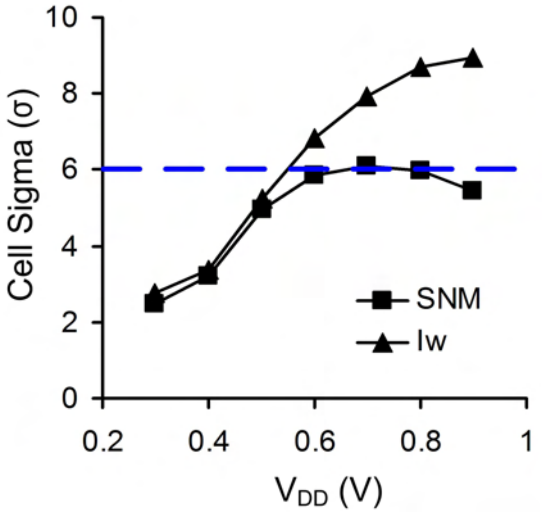

Figure 2.20. Read and write yield for a 65 nm node cell measured in terms of cell sigma, the minimum amount of variation necessary to cause a failure. Both yields decrease with V_{DD} scaling. For SNM, yield at high V_{DD} is limited by short channel effects.

metrics decrease. At high V_{DD} SNM yield saturates and even decreases due to the increasing effects of DIBL and a corresponding saturation in nominal SNM. Write-ability yield has a slightly stronger V_{DD} dependence. For a six-sigma yield, a $V_{DD} > 0.65$ is required, due to the constraint on SNM.

The speed of this approach enables SRAM designers to consider the effects on yield of new devices or layouts. So-called statistical or variation-aware design methodologies are indispensible for modern SRAM. To demonstrate how this model can inform SRAM design, the hypothetical 65nm node cell is modified so as to minimize the V_{DD} needed to achieve six sigma yield without increasing cell area.

There are limited options to effect tradeoffs in cell device strengths without impacting cell area. The specific options will be constrained by the supported process technology, which is not solely dependent on SRAM. For the sake of this example, small increases to the gate length are allowed ($<$ 6 nm), and the channel implant for the NMOS or PMOS devices can be adjusted on a global scale. This allows for five options: changing pull-down L, pass-gate L, pull-up L, NMOS V_{T0}, and PMOS V_{T0}. The sensitivities can be used to

Table 2.4. Sensitivities of SNM and I_W cell sigmas to L and V_{T0} changes

Design Option	Net Sensitivity Function	SNM Sensitivity (cell sigma / σ_x)	I_W Sensitivity (cell sigma / σ_x)	Net Effect (cell sigma / σ_x)
PD L	$\frac{\partial f}{\partial L_1} + \frac{\partial f}{\partial L_2}$	-0.072	0.059	0.013
PG L	$\frac{\partial f}{\partial L_3} + \frac{\partial f}{\partial L_4}$	0.500	-0.595	-0.095
PU L	$\frac{\partial f}{\partial L_5} + \frac{\partial f}{\partial L_6}$	-0.083	0.326	-0.243
NMOS V_{T0}	$\sum_{i=1}^{4} \frac{\partial f}{\partial V_{Ti}}$	0.440	-0.694	-0.254
PMOS V_{T0}	$\frac{\partial f}{\partial V_{T5}} + \frac{\partial f}{\partial V_{T6}}$	-0.243	0.308	-0.065

estimate the net effects on SNM and I_W cell sigma, as shown in Table 2.4. Since the cell is symmetrical, the effect of increasing **PD** L, for example, will be the sum of the effects from increasing L on each device. The net sensitivity function describes how the SNM and I_W sensitivities were calculated from device parameter sensitivities. A $V_{DD} = 0.6$ V was chosen because the cell sigma for both SNM and I_W are near the six sigma target at that bias.

One or more of these options can be used to balance SNM and I_W at $V_{DD} = 0.6$ V and thereby lower the six sigma V_{min}. The best option is that which has the greatest net increase to both metrics, which is determined by the sum of the sensitivities in the direction that balances SNM and I_W. For example, a decrease in **PD** L would improve SNM yield more than it would degrade I_W yield, with a net increase of 0.013 cell sigma per standard deviation of reduction in L. Decreases in gate length are not an allowed option in this example, though, and besides, the sensitivities are so small that large L changes would be needed.

The option with the next most positive net effect is a decrease in PMOS V_{T0}. This option degrades I_W slightly more than it improves SNM, but less so than the alternatives. The amount of V_{T0} change to balance SNM and I_W cell sigmas can be approximated as

$$\Delta V_{T0} = \frac{C_{\Sigma S} - C_{\Sigma W}}{\frac{\partial C_{\Sigma W}}{\partial V_{T0}} - \frac{\partial C_{\Sigma S}}{\partial V_{T0}}} \tag{2.23}$$

where $C_{\Sigma S}$ and $C_{\Sigma W}$ are the cell sigmas for SNM and I_W, respectively, and ignoring any second order effects. Using the numbers from Table 2.4, $\Delta V_{T0} = -1.77\sigma_{VT0} = -44$ mV is

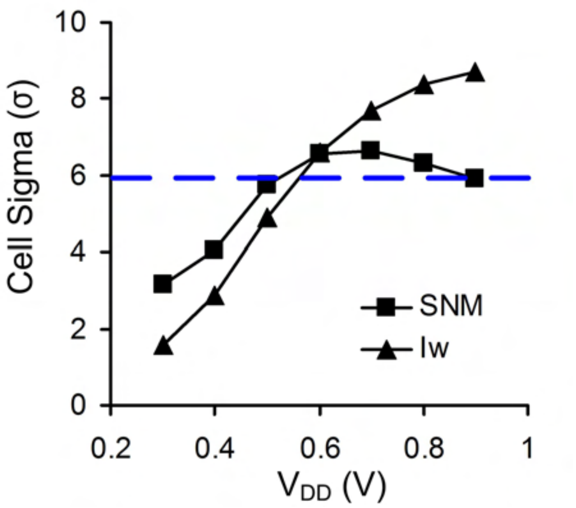

Figure 2.21. Using the sensitivities for the design options of Table 2.4, the **PU** V_{T0} is reduced to balance read and write cell sigma at $V_{DD} = 0.6$V. This enables six sigma yield at lower V_{DD}.

required. Fig. 2.21 illustrates read and write cell sigmas with the modified process. The cell sigmas are now balanced at $V_{DD} = 0.6$V. A six sigma yield is achievable at an even lower $V_{DD} = 0.58$V and is now constrained by I_W. The V_{DD} window for six sigma of SNM cell yield is widened too. This process can be repeated to optimize further the DC V_{min} at six sigma of variation.

It is important to remember that the model provides only a DC estimate. It is useful for the early stages of design, but transient simulations should be used to verify functionality and performance. A common technique is to use corner cases of variation in a transient simulation to check for yield. These corner cases are often simple and generic, taking a form such as a fast-NMOS / slow-PMOS corner, in which NMOS W and PMOS L and V_{T0} are increased by some fixed amount for all devices, and NMOS L and V_{T0} and PMOS W are decreased to match. At one sigma variation per (independent) parameter, the probability of this case is comparable to $\sqrt{18} \approx 4.2$ standard deviations of a single parameter Gaussian. It is too much variation to represent a one cell sigma design, yet it does not necessarily represent the worst cases of a four cell sigma design. A better approach is to use corner cases for transient simulations that correspond to the worst case variation vectors for SNM and I_W.

The recommended statistical SRAM design methodology therefore has two phases. The early phase should use a fast DC model to evaluate different technology options and optimize the cell design. In the later phase, once compact models can be developed and characterized, transient simulations should be used to verify performance and function at process corners determined from the DC worst case vectors.

2.5 NBTI and other reliability issues

Random dopant fluctuations and critical dimension variations are not the only challenges for SRAM reliability. Over time, transistor I-V curves can be affected by phenomena such as negative bias temperature instability (NBTI), hot carrier effects, or time dependent dielectric breakdown. Changes to the I-V curves from such phenomena affect SRAM metrics in the same manner as above.

NBTI is a particularly challenging problem, because it not only changes the strength of PMOS devices, it does so over time. In NBTI, a negative bias on the gate of a transistor attracts holes to the dielectric interface with the channel. A hole that interacts with a hydrogen-passivated dangling silicon bond may free the hydrogen ion, leaving behind a positive charge and increasing the absolute value of the PMOS threshold voltage. The distribution of charges on the interface is believed to be random and, in small devices, subject to discrete effects. It has been shown experimentally that the variance of threshold voltage due to NBTI varies inversely with the effective area of the device, in much the same way as it does for random dopant fluctuation [17]. This makes it difficult to screen out SRAM cells that function immediately after processing but which will eventually fail due to NBTI.

The statistical impact of NBTI on SRAM read stability is a subject of recent interest. Reddy *et al.* showed that NBTI could affect static noise margin (SNM) by as much as 8% at

$V_{DD} = 0.8$V, with an increasing effect as V_{DD} decreases [18]. La Rosa *et al.* extended this analysis to show that the variation in read stability increases as well, leading to larger failure counts in an array [19]. Ball *et al.* correlated V_{min} with the amount of NBTI measured in a single cell and showed consistent results for an array [20]. These results demonstrate that NBTI has a significant impact on cell reliability, yet several researchers have reported a relatively low sensitivity to PMOS threshold voltage in general [5, 21, 22], with Li *et al.* reporting a negligible sensitivity to NBTI specifically at $V_{DD} = 2.5$V [22]. Although such low sensitivities appear to contradict the results of [18, 19, 20, 23], the low sensitivities were reported at relatively high voltages of $V_{DD} \geq 0.9$V, and may underestimate the effects on low voltage metrics, such as V_{min}.

The mechanism of NBTI degradation on SRAM metrics such as V_{min} and SNM has been investigated with this model [24]. DC simulations of the SNM-constrained 65nm node design of Table 2.3 show that the sensitivity of SNM to PMOS V_{T0} increases at low voltages. As DIBL reduces the gains in SNM from increasing V_{DD}, the sensitivity of V_{min} to NBTI becomes comparable to those for other sources of variation. In addition, the sensitivity of V_{min} to NBTI is found to increase under certain combinations of variation in the SRAM NMOS devices. The most probable vector of parameter variations to set a given V_{min} for an array is identified and proposed as a useful corner case for transient simulations. This vector is shown to be dependent on NBTI, with increasing probability as mean NBTI increases.

For the purposes of understanding the sensitivity of V_{min} to NBTI, it is helpful to use SNM as an intermediary. $\partial SNM/\partial V_{T6}$ and $\partial SNM/\partial V_{DD}$ are easily extracted with the DC model. The sensitivity of V_{min} to NBTI can then be expressed as a combination of these sensitivities

$$\frac{\partial V_{min}}{\partial NBTI} = \frac{\partial SNM/\partial NBTI}{\partial SNM/\partial V_{DD}} \tag{2.24}$$

around the point where $V_{DD} = V_{min}$. Under a chip-wide measurement of NBTI, the V_{T0} of both PMOS transistors **PU5** and **PU6** will change. The net sensitivity of SNM to NBTI is the sum of $\partial SNM/\partial V_{T5}$ and $\partial SNM/\partial V_{T6}$, but since $\partial SNM/\partial V_{T6} >> \partial SNM/\partial V_{T5}$, only $\partial SNM/\partial V_{T6}$ is considered.

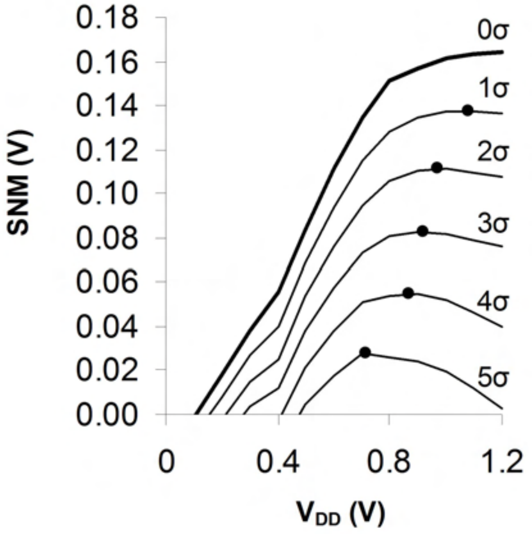

Figure 2.22. SNM vs. V_{DD} under total parameter variation ranging from zero (thick curve) to 5σ. Increases in DIBL reduce SNM at high voltages and shift the peak point (dot) toward lower V_{DD}.

In the discussion of Fig. 2.13, it was observed that V_{T0} sensitivities increased with lower V_{DD}. This is true for **PU6**, which determines the position of the upper shoulder of the butterfly curves at voltages comparable to V_{min}. The sensitivity to V_{T6} is still not a dominant one, but at 0.22 V/V it is approximately equivalent in magnitude to that to V_{T2} and almost half that to V_{T1} or V_{T3}. It is therefore essential that any investigation of NBTI in SRAMs consider operating voltage.

Not only does the sensitivity of SNM to NBTI increase at low V_{DD}, the effect on V_{min} is amplified by the sensitivity of SNM to V_{DD}. Fig. 2.22 illustrates the SNM vs. V_{DD} plot for an SRAM cell with different amounts of variation. The thick line represents the nominal design, with no variation. SNM tends to increase with V_{DD} nearly linearly at low voltages, until increases in DIBL diminish further gains. At low voltages of 0.5V - 0.6V, the slope of the curve is at its maximum value of 0.27 V/V. The sensitivity of V_{min} to NBTI can then be approximated following equation 2.24, resulting in relatively large sensitivities of approximately 0.8 V/V for the nominal cell design.

For an array-level V_{min}, the effect of W, L, and V_{T0} variations must be considered.

Table 2.5. Normalized 65nm SRAM variation vector

Device	PD1	PD2	PG3	PG4	PU5	PU6
W	-0.06	0.02	0.11	0.00	0.00	-0.03
L	0.26	-0.23	-0.59	0.00	0.00	0.08
V_{T0}	0.36	-0.36	-0.48	0.00	0.00	0.12

Since the V_{min} of an SRAM array is defined as the lowest operating V_{DD} at which all of the cells operate correctly, it is fairly straightforward to see that the array V_{min} will be equivalent to the highest cell V_{min} in that array.

In order to assess the impact of process variations, the SRAM cell was simulated under increasing amounts of variation along the vector specified in Table 2.5, which is the worst case vector for this cell at $V_{DD} = 0.7$V. Fig. 2.22 illustrates the degradation in SNM as a function of V_{DD}. At low voltages, increasing variation degrades SNM by a small amount (about 10 mV per sigma). At high V_{DD}, the increased degradation reflects a worsening of DIBL that is associated with the decrease in the length of PG3. As in the nominal case of no variation, an increase in DIBL counteracts the benefit to SNM of increasing V_{DD}, eventually causing SNM to decline with voltage under large amounts of variation. Fig. 2.22 illustrates that the peak SNM, a point where $\partial SNM/\partial V_{DD} = 0$, moves to lower voltages under increasing variation. At lower voltage the effect is less severe; however, the $\partial SNM/\partial V_{DD}$ sensitivity is still reduced. Under large amounts of variation, a cell that has low SNM and a low $\partial SNM/\partial V_{DD}$ sensitivity will exhibit greater sensitivity of V_{min} to phenomena such as NBTI.

Fig. 2.23 illustrates this increase in sensitivity over three sigma of variation. The DC V_{min} rises almost linearly with increasing variation, at the rate of 50 mV per sigma. At the same time, the sensitivity to NBTI increases by a factor of two. Referring back to equation 2.24, this increase can be attributed mostly to a decrease in the sensitivity of SNM to V_{DD} (-30% at 0.5 V). There is a modest increase in $\partial SNM/\partial V_{T6}$ of less than 10% over this range of variation. These results are consistent with the reported measurements of $\partial V_{min}/\partial NBTI$ and the observation that cells with large NBTI sensitivities

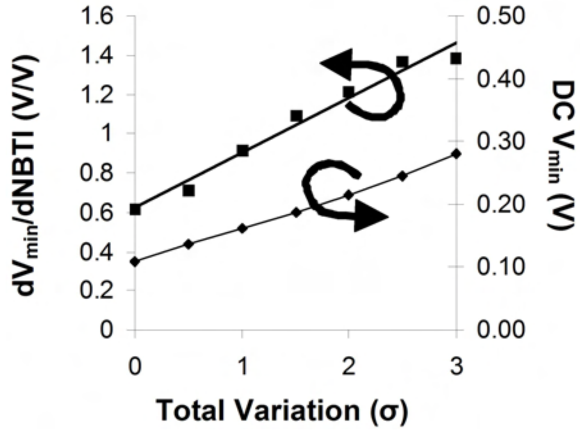

Figure 2.23. Sensitivity of V_{min} to NBTI and DC V_{min} vs. total variation along the vector in Table 2.5. The sensitivity increases with total variation even though the majority of this variation is in the NMOS devices. The DC V_{min} also rises with total variation, suggesting that array V_{min} will be larger than the V_{min} of a nominal cell and confirming the importance of corner case simulation for accurate V_{min} prediction.

exhibit large amounts of parameter variations in other devices, especially the pass-gates [20]. It is therefore important to consider NBTI as a source of possibly significant V_{min} degradation, not only for the increased sensitivity of low voltage metrics, but also for the increased sensitivity caused by other sources of variation.

For the purpose of statistical design, it would be of great value to predict V_{min} for an array of a certain size, given a cell design and knowledge of the statistics of parametric variations, but unfortunately this is a very challenging problem. For example, a 32Mb SRAM array has a 3% chance of including at least one bitcell with at least six sigma of total variation. The worst case vector at six sigma of variation has a V_{min} of 0.7V; however, this is just one of many vectors with six sigma of total variation, and the probability of its occurrence is infinitesimal. The best solution one could achieve is a probability distribution for V_{min}, dependent on the V_{min} calculated for each variation vector and the corresponding probability of that vector; however, the computation time would be prohibitive. For this reason, the worst case vectors for SNM were used, in order to show generally how the V_{min} of an array is likely to exhibit a large sensitivity to NBTI. The precise sensitivity will depend on the distribution of variations in that particular array, but the trends can be used to guide design.

In order to apply to SRAM cells after burn-in, the worst case vector should be modified to include the effects of NBTI. This changes the worst case vector in two ways: it decreases the total variation needed to cause a failure (cell sigma), and it increases the relative contribution of the PMOS transistors. To model the yield, two additional dimensions of variation are added to represent the V_T shift caused by NBTI in each of the PMOS devices, **PU5** and **PU6**. For NBTI, the relation between mean and standard deviation reported in [17] is used:

$$\sigma_{\Delta V_T} = \frac{1}{2}\sqrt{\frac{K_1 K_0 T_{ox} \mu_{\Delta V_T}}{2WL}} \tag{2.25}$$

where $K_0 = 2/\epsilon_{ox}$, $K_1 = 2.7$ is a constant determined experimentally from long-channel devices, T_{ox} is the effective oxide thickness with ϵ_{ox} dielectric constant, and a factor of $1/\sqrt{2}$ is used to adapt the equation to a single device from a mismatch calculation. The ratio of the variance of ΔV_T from NBTI to σ^2_{VT0} from random dopant fluctuation was shown to be proportional to mean NBTI and invariant to device size. A factor of 1/2 was added to equation 2.25 to match this relationship.

Fig. 2.24 illustrates the composition of the worst case vector for SNM at $V_{DD} = 0.3$V and 0.6V. At $V_{DD} = 0.3$V, the inclusion of NBTI decreases the amount of NMOS V_T variation needed to cause a failure, thereby increasing the relative contribution of PMOS V_T from 15% to 36% of the worst case vector. Similarly, for $V_{DD} = 0.6$V, NBTI shifts the worst case vector to one with less total variation and a greater contribution of the PMOS V_T. A corner case simulation neglecting NBTI or PMOS variations could therefore significantly underestimate V_{min}.

By decreasing the contributions of NMOS V_T to the worst case vector, the inclusion of NBTI also decreases the minimum total variation necessary to cause a failure, increasing the probability of occurrence in an array (Fig. 2.25). The magnitude of the worst case vectors (cell sigma) is reported as a function of V_{DD}. This corresponds to the minimum amount of variation that could set V_{min}, and it represents the most probable combination of

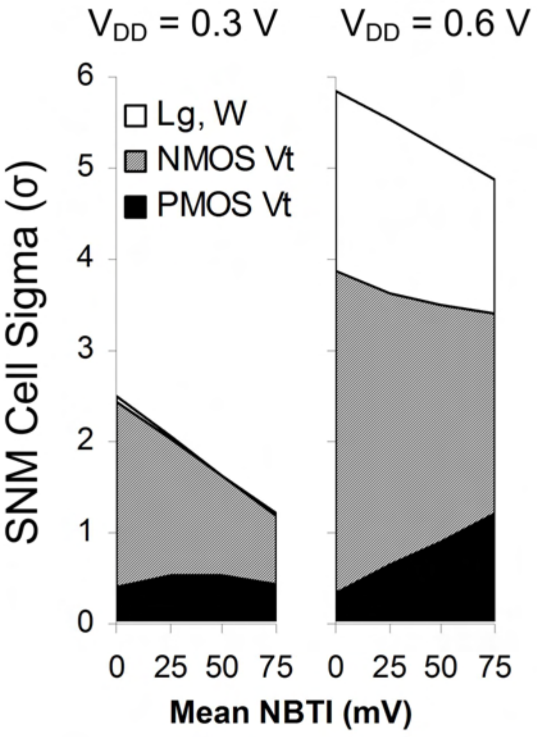

Figure 2.24. Relative contributions of variation sources such as NMOS and PMOS V_T in the worst case vector for SNM as a function of increasing NBTI. Increasing NBTI shifts the worst case vector to a more probable combination (less total variation) primarily by decreasing the contribution from NMOS V_T. For V_{DD} = 0.3V, the worst case vector consists almost entirely of variations among the NMOS and PMOS V_T. At higher voltages, the contribution from other sources increases, due mainly to DIBL.

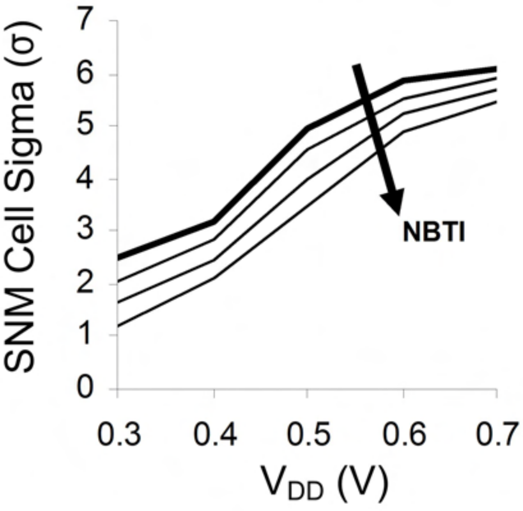

Figure 2.25. Cell sigma decreases at all V_{DD} for mean NBTI of 0, 25, 50, and 75mV. As an additional source of variation, NBTI decreases the amount of variation needed in other device parameters to set V_{min}.

parameter variations to set the V_{min} at a given V_{DD}. Cell sigma decreases with the addition of NBTI at a rate of about 1.6 sigma / 100 mV. This is a significant decrease for cells that are typically designed for six sigma of yield, and it shows that a statistical SRAM design ought to include NBTI as a source of variation.

2.6 Conclusion

An understanding of the origins and mechanisms of variation in SRAM is crucial for modern designs. In this section, a fast and accurate model for DC metrics was developed. This model uses a handful of targets to analytically represent the I-V behavior of short-channel MOSFETs, even in the presence of minor physical inaccuracies in the underlying equations. The model has been validated to several hundred fabricated cells among different process, technology, bias, and temperature conditions, as well as to TAURUS device and mixed-mode simulations.

From nominal simulations of device metrics, sensitivities can be extracted. These sensitivities can be used to understand the relative importance of variations on different parameters in different devices. They enable a much faster projection of cell yield than is achievable with a pass/fail metric. By using the sensitivities to quantify design tradeoffs among device and circuit options, a cell can be designed for high yield in both read and write function at low V_{DD}. The statistical design of SRAM should consist of two phases: a DC phase for optimization and a transient simulation using DC corner cases to ensure high yield.

Modeling also elucidates the mechanisms of time-dependent reliability challenges, such as NBTI. The sensitivity of SNM to NBTI was shown to increase for low V_{DD} and with parameter variation in both NMOS and PMOS devices, to a level that is comparable with NMOS parameter variations. Future analyses of NBTI should be sure to consider the effects of operating voltage and variation. The statistical design methodology presented here can be easily adapted for NBTI by including it as an additional device parameter.

Similar analyses can be performed for other reliability phenomena or to evaluate options for future technology nodes, such as alternative device structures or processes. In particular, high SNM and I_W sensitivities to gate length and threshold voltage encourage research into technology options that reduce their variability. The DC model in this work can provide fast and accurate insight into the advantages and tradeoffs of such options.

2.7 References

[1] A. Dixit, K. G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, and K. De Meyer. Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: Role of line-edge-roughness. *IEEE International Electron Devices Meeting*, pages 709–712, 2006.

[2] Y. Okayama, T. Saito, K. Nakajima, S. Taniguchi, T. Ono, K. Nakayama, R. Watanabe, A. Oishi, A. Eiho, T. Komoda, T. Kimura, M. Hamaguchi, Y. Takegawa, T. Aoyama, T. Iinuma, K. Fukasaku, R. Morimoto, K. Oshima, K. Oono, M. Saito, M. Iwai, N. Nagashima, and F. Matsuoka. Suppression effects of threshold voltage variation with Ni FUSI gate electrode for 45nm node and beyond LSTP and SRAM devices. *IEEE Symposium on VLSI Technology*, pages 96–97, 2006.

[3] E. Seevinck, F. List, and J. Lohstroh. Static-noise margin analysis of MOS SRAM cells. *IEEE Journal of Solid-State Circuits*, pages 748–754, 1987.

[4] T. Ichikawa and M. Sasaki. A new analytical model of SRAM cell stability in low-voltage operation. *IEEE Transactions on Electron Devices*, pages 54–61, 1996.

[5] B. Calhoun and A. Chandrakasan. Static noise margin variation for sub-threshold SRAM in 65-nm CMOS. *IEEE Journal of Solid-State Circuits*, pages 1673–1679, 2006.

[6] Q. Chen, A. Guha, and K. Roy. An accurate analytical SNM modeling technique for SRAMs based on butterworth filter function. *IEEE International Conference on VLSI Design*, pages 615–620, 2007.

[7] S. Wolf. *Silicon processing for the VLSI era Volume 3: The submicron MOSFET*. Lattice Press, 1995.

[8] A. Kaneko, A. Yagishita, K. Yahashi, T. Kubota, M. Omura, K. Matsuo, I. Mizushima, K. Okano, H. Kawasaki, T. Izumida, T. Kanemura, N. Aoki, A. Kinoshita, J. Koga, S. Inaba, K. Ishimaru, Y. Toyoshima, H. Ishiuchi, K. Suguro, K. Eguchi, and Y. Tsunashima. High-performance FinFET with dopant-segregated schottky source/drain. *IEEE International Electron Devices Meeting*, pages 893–896, 2006.

[9] M. Agostinelli, J. Hicks, J. Xu, B. Woolery, K. Mistry, K. Zhang, S. Jacobs, J. Jopling, W. Yang, B. Lee, T. Raz, M. Mehalel, P. Kolar, Y. Wang, J. Sandford, D. Pivin, C. Peterson, M. DiBattista, S. Pae, M. Jones, S. Johnson, and G. Subramanian. Erratic fluctuations of SRAM cache vmin at the 90nm process technology node. *IEEE International Electron Devices Meeting*, pages 655–658, 2005.

[10] C. Wann, R. Wong, D. Frankt, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright, and J. Sudijono. SRAM cell design for stability methodology. *IEEE VLSI-TSA International Symposium*, pages 21–22, 2005.

[11] A. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin. Fluctuation limits and scaling opportunities for CMOS SRAM cells. *IEEE International Electron Devices Meeting*, pages 659–662, 2005.

[12] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene. Read stability and write-ability analysis of SRAM cells for nanometer technologies. *IEEE Journal of Solid-State Circuits*, pages 2577–2588, 2006.

[13] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake. Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit. *International Solid State Circuits Conference*, page 34.5, 2006.

[14] R. Wong. Cmos sram cell with pfet passgate devices. *United States Patent No. 6341083*, 2002.

[15] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr. A 153MB-SRAM design with dynamic stability enhancement and leakage reduction in 45nm high-κ metal-gate CMOS technology. *International Solid-State Circuits Conference*, pages 376–377, 2008.

[16] D. Burnett. Statistical design issues of SRAM bitcells and sense amps. *IEEE Silicon on Insulator Conference*, 2006. Short Course.

[17] S. Rauch. The statistics of NBTI-induced v_t and β mismatch shifts in pMOSFETs. *IEEE Trans. Device and Materials Reliability*, pages 89–93, 2002.

[18] V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan. Impact of negative bias temperature instability on digital circuit reliability. *Intl. Reliability Physics Symp.*, pages 248–254, 2002.

[19] G. La Rosa, W. L. Ng, S. Rauch, R. Wong, and J. Sudijono. Impact of NBTI induced statistical variation to SRAM cell stability. *Intl. Reliability Physics Symp.*, pages 274–282, 2006.

[20] M. Ball, J. Rosal, R. McKee, WK Loh, T. Houston, R. Garcia, J. Raval, D. Li, R. Hollingsworth, R. Gury, R. Eklund, J. Vaccani, B. Castellano, F. Piacibello, S. Ashburn, A. Tsao, A. Krishnan, J. Ondrusek, and T. Anderson. A screening methodology for vmin drift in SRAM arrays with application to sub-65nm nodes. *IEEE International Electron Devices Meeting*, pages 705–708, 2006.

[21] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, pages 1859–1880, 2005.

[22] X. Li, J. Qin, B. Huang, X. Zhang, and J. B. Bernstein. SRAM circuit-failure modeling and reliability simulation with SPICE. *IEEE Trans. Device and Materials Reliability*, pages 235–246, 2006.

[23] S. V. Kumar, K. H. Kumar, and S. S. Sapatnekar. Impact of NBTI on SRAM read stability and design for reliability. *International Symposium on Quality Electronic Design*, 2006.

[24] A. Carlson. Mechanism of increase in SRAM V_{min} due to negative bias temperature instability. *IEEE Trans. on Device and Materials Reliability*, pages 479–487, 2007.

Chapter 3

Device Techniques for Reducing Variation in SRAM

3.1 Introduction

An accurate model provides insights into the causes of SRAM variation and the ways it can be reduced. Sensitivity analyses for 6-T cells exhibit a few characteristic trends, irrespective of the specific technology node or process of the cell. Specifically, both read stability and write-ability metrics are highly sensitive to the pass-gate device parameters, making the control of those transistors critical. Within a device, the sensitivity to device threshold voltage is generally greatest, followed by gate length and then channel width. By designing for low device variation, SRAM variation can also be reduced.

In Section 2.4, a method for optimizing an SRAM bitcell design was demonstrated. This method approximated the tradeoffs associated with device parameters by considering their first order effects on drive current. For example, increasing the gate length of a short channel MOSFET will increase V_{TSAT}, thereby reducing drive current. Such a change can also have second order effects on the variability of the device parameters themselves. Ignoring any reverse short channel effects, the sensitivity to gate length variations $\partial V_{TSAT}/\partial L$ will also

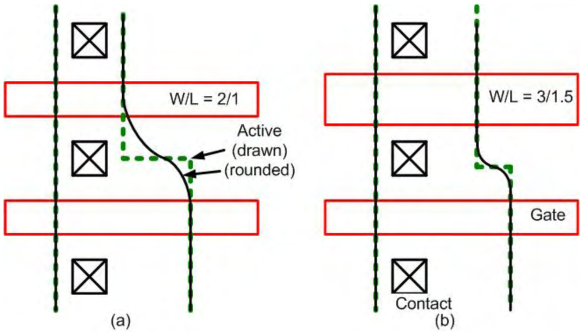

Figure 3.1. Different layouts can be used to achieve the same device sizing ratio. Layouts with large steps in width between adjacent devices are more susceptible to corner rounding (a). From a variability perspective, it is preferable to achieve the device sizing with a longer L_G (b). Not only is the amount of corner rounding reduced, but V_{T0} variation (proportional to $1/\sqrt{WL}$) is reduced as well.

decrease. The magnitude of SNM and I_W sensitivities to L will therefore decrease as well, reducing the effect on yield of further L increases. In addition, the standard deviation of the linear threshold voltage σ_{VT0} in a uniformly doped channel follows [1]:

$$\sigma_{VT0} \propto N_a^{1/4} \left(W_{eff} L_{eff}\right)^{-1/2} \tag{3.1}$$

where N_a is the average channel doping, W_{eff} is the effective channel width, and L_{eff} is the effective gate length. As L increases, σ_{VT0} decreases. A similar dependence can be found for devices with corner rounding in the active layer, as illustrated in Fig. 3.1. In such a device, the width variation σ_W increases with the magnitude of the step in the active layer. Such second order effects are small in comparison to the first order changes to nominal SNM or I_W. Nevertheless, these effects favor pass-gate devices with less aggressive W and L scaling for modern SRAM designs.

More importantly, they suggest a direction for reducing SRAM variability: designing layouts or processes to reduce variation at the parameter level, e.g. σ_W, σ_L, or σ_{VT0}. Among these parameters, sensitivity to V_{T0} is usually dominant. Fig. 3.2 illustrates the relative importance of V_{T0} variation for the 65nm node cell investigated in Section 2.4. To generate the data in Fig. 3.2, the most probable point of failure was simulated for different V_{DD}, and the component variations were separated by parameter type, W, L, or V_{T0}. The relative contribution for each type was calculated using the sum of the squares

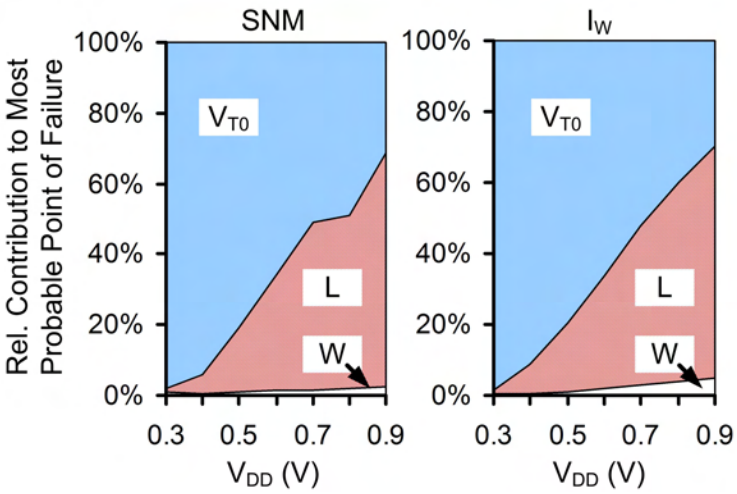

Figure 3.2. Variations in V_{T0} make up the dominant contribution to the most probable point of failure at low V_{DD}, with increasing contributions from L variations as V_{DD} increases.

of the variations among all devices. V_{T0} variations are most important for yield at low V_{DD}. As V_{DD} increases, the relative contribution of L variations becomes more significant and eventually greater for $V_{DD} > 1.0$ V. Width variations are the least significant, in part due to a low assumed σ_W. Techniques for reducing V_{T0} and L variations are therefore of particular interest for improving SRAM yield.

3.2 Reducing variation from threshold voltage

In planar bulk or partially-depleted silicon-on-insulator (PDSOI) MOSFETs, random dopant fluctuations (RDF) are currently the dominant source of σ_{VT} and are expected to remain so while $L_G > 20$nm [2]. RDF is an intrinsic problem that arises from discretization of the number and placement of dopant atoms. The only ways to reduce the effects of random dopant fluctuations are to reduce the sensitivity of V_{T0} to doping or to eliminate channel dopants entirely.

Classically, V_T is determined by the sum of voltage drops from the gate to the source at the threshold of inversion. In uniformly doped channels,

$$V_T = \Phi_{MS} + 2\Phi_B + \frac{2\sqrt{q\epsilon_{si}N_a\Phi_B}}{C_{ox}} \tag{3.2}$$

where Φ_{MS} is the difference in the work functions of the gate and channel regions, $2\Phi_B$ is the band-bending in the silicon, q is the electron charge, ϵ_{si} is the dielectric constant of the silicon channel, and C_{ox} is the gate oxide capacitance per unit area. Insofar as Φ_B depends on N_a, each term of the V_T equation depends on dopants; however, the first two terms are logarithmic functions of N_a and exhibit a relatively weak dependence. The sensitivity of V_T to N_a can be approximated by differentiating the final term:

$$\frac{\partial V_T}{\partial N_a} \approx \frac{\sqrt{q\epsilon_{si}\Phi_B}}{C_{ox}\sqrt{N_a}} \tag{3.3}$$

which decreases with large C_{ox} or high N_a. Increases in C_{ox} are generally limited by constraints on gate leakage current, which increases exponentially with decreasing oxide thickness. Even with the introduction of high-k dielectrics, C_{ox} is not expected to increase more than a factor of two in high performance devices. The remaining option is a very high N_a. High doping concentrations are also necessary to control short channel effects and block punchthrough.

There are significant disadvantages to increasing doping concentrations, though. Ionized dopants degrade carrier mobilities via scattering, lowering drive currents. Very high dopant concentrations reduce the depletion width at the drain and source junctions, increasing junction capacitance and off-state current via tunneling mechanisms such as gate-induced-drain-leakage (GIDL). High doping concentrations also increase V_T, reducing drive current. Advances in channel doping profiles have enabled dopant concentrations on the order of 10^{19}cm^{-3} in modern devices, with continued increases unlikely to be practical. Above all, the variation in dopant number increases as $\sqrt{N_a}$, resulting in a net increase in σ_{VT0} as given by Eqn. 3.1 [1]. There is little room to continue increasing N_a or C_{ox} in future technology generations. The solution to random dopant fluctuation is not to be found in continued planar MOSFET scaling.

3.2.1 Alternative device architectures

Alternative MOSFET structures (Fig. 3.3) have been proposed for SRAM in which V_T control can be achieved without the use of channel dopants, thereby greatly reducing device sensitivity to random dopant fluctuation. Such architectures include fully depleted silicon-on-insulator (FDSOI) [3], FinFETs (double gate) [4], triple-gate [5], and gate-all-around devices [6]. Although all of these device architectures provide improved scalability relative to current planar bulk or partially-depleted SOI technologies, the transition to a new architecture has been continually put off in favor of incremental enhancements such as strained channels or, most recently, high-k gate dielectrics. In part this reflects the enormity of the investment and risk associated with developing the design infrastructure necessary for a new device architecture. In addition, careful process and layout optimizations have allowed for tolerable, though decreasing, array yields. However, as array size continues to grow and V_T variation continues to increase, the transition to a new architecture may be inevitable if SRAM scaling is to continue. Maintaining good yield will require a device architecture in which V_T is set by parameters with relatively low variability, such as the physical dimensions of the channel and the work function of the gate metal.

FDSOI

FDSOI (Fig. 3.3a) is the most planar of these architectures. In FDSOI, the depletion region extends throughout the thickness of the channel layer. Scaled FDSOI designs can eliminate channel dopants, enabling higher carrier mobilities and reducing drain-to-body capacitance, which provide for improved circuit performance with lower dynamic power consumption. Devices with undoped channels have negligible depletion charge and capacitance, which yields a steep subthreshold slope. Most importantly, the absence of channel dopants all but eliminates the V_{T0} variation due to random dopant fluctuation (errant source / drain dopants may cause a small residual RDF). The V_T is then a function of the gate work function and the thickness of the silicon channel layer, t_{si}. Variations in t_{si} tend to be either small ($\sigma_t < 5$ Å, due to roughness [7]) or of a common mode. As a planar,

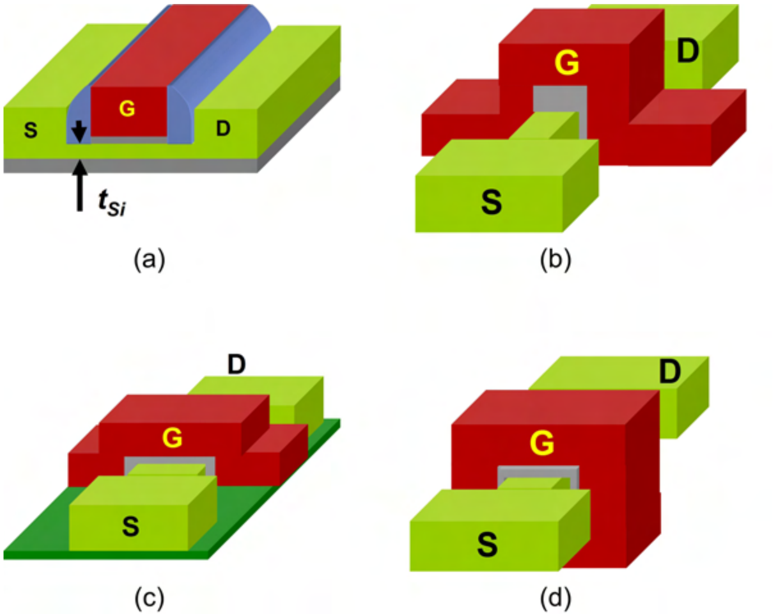

Figure 3.3. Alternative device architectures do not rely on dopants for V_T control, and can therefore use undoped channels. A significant advantage of such architectures is a robust V_T control. The scalability of each architecture improves with the number of sides under gate control. FDSOI (a) is the least scalable. Double gate (b) and triple gate (c) architectures have improved scalability while remaining manufacturable. Gate-all-around architectures (d) offer the ultimate gate control but are very difficult to manufacture.

single gate technology, FDSOI can accommodate a wide and continuous range of device widths, enabling ideal beta ratios in SRAM designs. Existing bulk designs could be ported to FDSOI with the least amount of design effort, relative to other device architectures.

The problem with FDSOI is its scalability. Silicon film thicknesses of $t_{si} < L_G/4$ are needed for good short channel behavior [8]. In addition to being expensive to manufacture uniformly, channel thicknesses smaller than a few nanometers are expected to have degraded on-state currents due to quantum confinement effects and increased parasitic resistance. These effects will make it difficult to scale FDSOI much beyond the 22nm node.

Double gate

Double gate architectures, such as FinFETs (Fig. 3.3b), can exhibit the same short channel control with a relaxed body thickness of $t_{si} < 2L_G/3$ [9]. FinFET devices enjoy similar improvements to FDSOI in carrier mobility and subthreshold slope when an undoped channel is used. The vertical fin of the FinFET can be manufactured with conventional lithography and anisotropic etching processes. As with FDSOI, the V_T is set by the gate work function and the silicon thickness. The thickness control from lithography will be comparable to that for CD control ($\sigma_L > 3$ nm) and higher than the thickness variation in FDSOI. Although low V_Ts are difficult to achieve simultaneously in NMOS and PMOS logic devices, a single, mid-gap gate work function can be used for high V_T applications, such as SRAM. In addition, FinFETs have a lower parasitic device capacitance because both depletion and junction capacitances are effectively eliminated, which reduces the **BL** capacitive load.

The disadvantages of FinFETs are in manufacturing and circuit design. The added height of the FinFET requires a lithography system with a higher depth of focus in order to minimize gate length variation. The tall aspect ratio can also present problems for implant, etching, and planarization processes. Achieving different gate work functions for NMOS and PMOS or between logic and SRAM is an additional challenge. Also, the reliability of such devices at a very large integration scale is presently unknown. For circuit design, layouts are confined to quantized widths in units of individual fins. Because the optimal beta ratios in modern SRAMs are closer to 1.5, cell designs must incur either a high area penalty from using multiple fin devices or a penalty to nominal SNM from using single fin **PD** and **PG** devices. In spite of these challenges, FinFETs are widely expected to improve SRAM yield. FinFET-based SRAMs have already been demonstrated in silicon with excellent nominal SNM and leakage control [10, 11, 12]. They are also expected to improve array yield, even when minimum width devices are used for the **PD** transistors [13].

A unique advantage of FinFETs is that the front and back gates can be separated and independently controlled. Independent gate operation is achieved by selectively removing the gate material directly on top of the fin, leaving the gates electrically isolated [14]. This enables several new cell designs, which can be used to further reduce variability. In addition to enabling additional connectivity within the SRAM cell, it provides an alternative direction for future technology development beyond gate length scaling. Several independently-gated FinFET SRAM designs have demonstrated improved performance and yield, and will be examined in detail in Section 4.4. FinFETs therefore are a promising device architecture for continued SRAM scaling, due to both a robust V_T control and the opportunity for further enhancements with independent gating.

Triple gate

Triple gate devices are like FinFETs, except that the gate also controls the top surface of the channel region. Subthreshold current is suppressed throughout the channel, except for the region at the base, farthest away from the gate. There are several variations on triple gate architectures that vary on their control of this current [5, 15, 16, 17]. Fossum *et al.* have shown that without some kind of subthreshold current control from the bottom surface, the triple gate device must be either tall and narrow or short and wide [18]. In other words, it must resemble either a FinFET or a planar FDSOI device. The use of a ground plane with a bulk triple gate device improves short channel control by limiting leakage current at the bottom surface [17]. This improves the scalability of the structure with a relaxed body thickness requirement of almost $t_{si} < L_G$.

Triple gate devices may be easier to manufacture than FinFETs by performing a slight recess of the shallow trench isolation (STI) in a planar process. The STI recess is a timed etch process, but it can be well controlled by using a shallow implantation to selectively increase the etch rate to a specific depth [19]. The gate stack is fabricated over the corners and sides of the exposed active area.

V_T variation stems from several small sources. Triple gate devices are more susceptible to width variations than planar transistors but less so than FinFETs. Variation in the height can also affect V_T, particularly if there is an underlap between the STI recess depth and the ground plane depth. To minimize this variability, the device structure must be designed with a slight overlap, corresponding to a deeper STI recess. The overlap will increase gate capacitance and may also increase gate leakage current if the gate oxide is too thin. Corner rounding can also affect V_T variation, but the effect is minimized if the channel region is undoped. Finally, random dopant fluctuations from the ground plane will impact V_T, but at a fraction (W_{layout}/W_{eff}) of that for the planar bulk MOSFET. Overall, V_T variation is expected to be much less than that for comparably sized planar devices, making the bulk triple gate architecture worthy of further analysis for SRAM.

Gate-all-around

Further improvements in short channel control can be achieved with gate-all-around devices. Gate-all-around enables even more relaxed channel dimensions (W or $t_{si} < L_G$), but the improvements are marginal. The devices are expected to exhibit minimum V_T variation, but again with only marginal improvements over FinFET or triple gate devices. Table 3.1 shows the silicon thickness constraints and approximate V_T sensitivities to variations in device dimensions. The major disadvantage of gate-all-around architectures is that they are very expensive and difficult to manufacture. It is therefore unlikely that they will be adopted for SRAM.

Table 3.1. Projected scaling and variability of alternative device architectures at W = L = 30nm (estimated from [20, 15])

	FDSOI	FinFET	Triple Gate	Gate-All-Around
Thickness Requirement	$t_{si} < L/4$	$t_{si} < 2L/3$	$t_{si} < L$	$t_{si} < L$
$\partial V_{T0}/\partial t_{si}$	11 mV/nm	5 mV/nm	1.5 mV/nm	0.5 mV/nm
$\partial V_{T0}/\partial W$	3 mV/nm	3 mV/nm	1.5 mV/nm	0.5 mV/nm

3.2.2 Triple Gate Bulk SRAM

To address the scaling challenges associated with random dopant fluctuation, it is likely that either a FinFET or triple-gate device architecture will be adopted. Of the two, triple-gate devices are the most similar to current planar devices. They offer the best layout efficiency with a low aspect ratio that provides for high on-currents and relatively easy manufacturing. Since it is more of an incremental technology step, triple-gate devices are more likely to be implemented first.

In this section, the characteristics of a triple-gate SRAM cell are compared against those of a planar cell [21]. Given the myriad tradeoffs present in SRAM design, it is important to compare different device technologies across multiple cell metrics, including cell area, read stability, write-ability, access speed, and yield. It is not sufficient to claim a cell as superior based on area and nominal SNM alone. Instead, fair comparisons must be made at multiple design points to investigate the tradeoffs associated with each device technology.

For an initial comparison, triple-gate and planar devices are designed using 3-D device simulations [22]. Both devices employ an equivalent super-steep-retrograde (SSR) dopant profile in the channel region, with a peak dose of $2 \times 10^{19}\text{cm}^{-3}$ at a depth of 15nm and a gradient of 4nm / decade [23]. In the planar device, this profile represents the well doping, while in the triple-gate device it defines the ground plane. Gate work functions are adjusted such that $V_{T0,NMOS} = 0.34$V and $V_{T0,PMOS} = 0.37V$ at 100nA of current. Minimum-width devices are used with $W_{layout} = L_G = 20$nm. Same layout configurations (and thus layout areas) are assumed. The triple-gate devices have a nominal height of 14nm, resulting in a larger effective channel width of $W_{eff} = W_{layout} + 2 \times 14\text{nm} = 48\text{nm}$ (Fig. 3.4). The larger

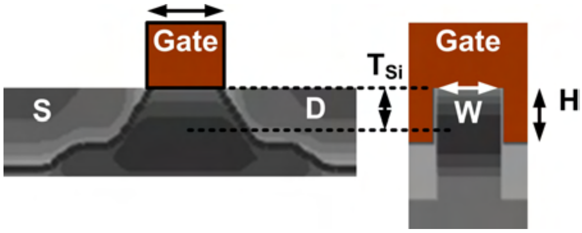

Figure 3.4. The triple-gate bulk device conducts current along the top and sidewalls of the channel, using a ground plane to control short channel effects. The improved gate control provides for increased robustness to random dopant fluctuations.

W_{eff} allows for a higher I_{DSAT}, while the multi-gate control provides lower subthreshold swing, DIBL, off-state current, and body effect.

The effect of variations in device width, length, and height (triple-gate only) are investigated with 3-D simulations of variations in individual parameters up to several standard deviations. Combinations of variations (e.g. W and L_G) are treated as affecting the I-V targets independently. (This assumption was verified with a few representative simulations.) Although standard deviations in device dimensions are highly dependent on the patterning process, a representative standard deviation of $\sigma_W = \sigma_L = 2.0$nm is used for a ballpark analysis. Triple-gate devices have a stronger narrow-width effect but a lower V_T rolloff effect than planar devices. The triple-gate device was found to be relatively insensitive to height variations beyond a certain depth, where the gate overlaps with the ground plane. On the other hand, shallow devices with severe gate-to-ground plane underlaps exhibit very high sensitivities to height variations. The optimal triple-gate design should therefore have a small amount of overlap to mitigate this sensitivity. In the following analysis, a small overlap is assumed so that height variations are negligible.

Variations in V_{T0} are determined using atomistic, Monte Carlo 3-D simulations from the nominal device dimensions [24]. A set of 100 NMOS and PMOS devices yields Gaussian distributions for V_{T0} with standard deviations in the planar devices of $\sigma_{VT0,NMOS} = 27$mV and $\sigma_{VT0,PMOS} = 30$mV. In the triple-gate devices, improved gate control of the channel results in smaller variations, with $\sigma_{VT0,NMOS} = 10$mV and $\sigma_{VT0,PMOS} = 12$mV. The ratio of planar to triple-gate σ_{VT0} is nearly identical to the inverse ratio of effective widths.

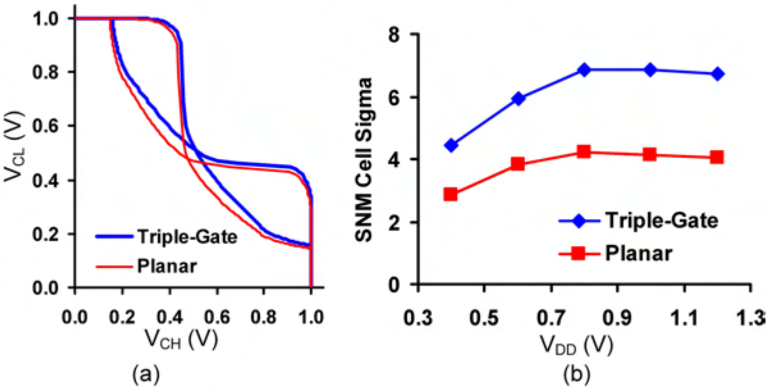

Figure 3.5. Even though nominal SNM is comparable between the triple-gate and planar SRAM cells with same V_{T0} and cell area (a), the read yield for the triple-gate cell is much higher. The increase can be attributed primarily to lower σ_{VT0}.

I-V targets are extracted from cases near the +/- 3 sigma points in the distribution to capture any short-channel and other effects due to the change in dopant profile. Because discretization effects shift the means of the V_{T0} distributions [25, 24], the I-V targets from simulating an average-V_{T0} device with a discretized profile differ from those with a continuous profile. The I-V targets from the discretized profile are therefore scaled by a factor of $I_{nom,cont}/I_{nom,disc}$ for compatibility, where $I_{nom,cont}$ is the I-V target of a nominal device with a continuous dopant profile. $I_{nom,disc}$ is the I-V target from the device closest to the average V_{TLIN} from the distribution of Monte Carlo simulations with discrete dopant profiles. A similar adjustment is made for V_{TLIN} and V_{TSAT} targets; however, an additive shift is made rather than a multiplicative one.

Because the nominal threshold voltages are equal in the two designs, the nominal read stability and write-ability curves are similar. The SNM of the nominal triple-gate and planar designs are almost equal (179mV vs. 186mV at $V_{DD} = 1.0$V), as illustrated in Fig. 3.5a. The SNM of the planar cell is slightly higher due to the increased body effect in the pass-gate devices; however, the triple-gate cell exhibits approximately two sigma higher read yield over a wide range of V_{DD} (Fig. 3.5b). For write-ability, I_W in the triple-gate device is much higher, owing to the increased drive strength of the triple-gate devices. Normalized to NMOS I_{DSAT} for each device technology, the shape of the curves is very similar (Fig. 3.6a). Write yield is significantly higher for the triple-gate cell, due primarily to the reduced

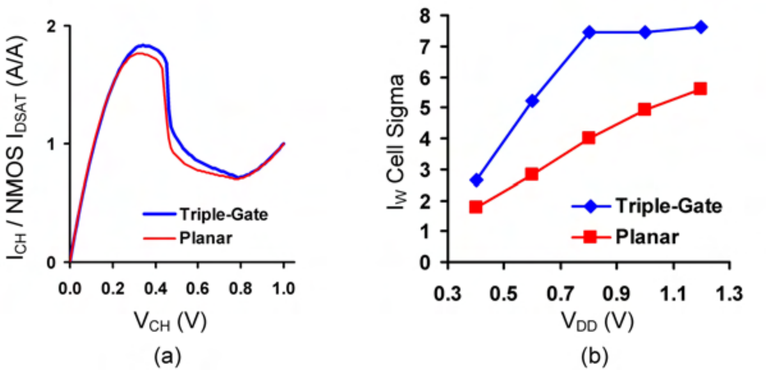

Figure 3.6. Write-ability current in the triple-gate cell is significantly higher than in the planar cell, due to the increased drive strength of the triple-gate **PG** device. Normalized to **PG** I_{DSAT}, the curves are actually very similar (a). As for read stability, the write yield is significantly higher due to low σ_{VT0} and V_T rolloff effects (b).

PMOS σ_{VT0} and V_T rolloff effects. A similar amount of PMOS V_T shift can prevent the write in both planar and triple-gate cells, but because it is less probable to occur in the triple-gate cell, the statistical yield is higher. The write yield increases faster as a function of V_{DD} for the same reason, and saturates at a high level of 7.5σ corresponding to the amount of variation needed to cause a conductive short in the PMOS.

The triple-gate cell has much higher read and write yields than the planar cell using minimum-sized devices; however, optimized SRAM designs typically use wider **PD** transistors to improve read yield, with a beta ratio around 1.5 or 2. For a planar cell, **PD** sizing involves a tradeoff between cell area and SNM (and, to a much lesser extent, write access speed). Increases in the **PD** width directly result in a corresponding increase in cell area. For a cell with average active and gate pitches of P, the cell area increases linearly with W_{PD}:

$$A = 2P \times [4P + 2(W_{PD} - W_{nom})] \tag{3.4}$$

Since SNM increases monotonically with W_{PD} but with diminshing returns, there is an optimal point for maximizing the SNM to cell area ratio. Fig. 3.7a illustrates this curve for a planar cell with $P = 90$nm, normalized to the SNM/area ratio at $W_{PD} = W_{min}$, for several V_{DD}. At $V_{DD} \geq 1.0$V, the optimal point is at a beta ratio of 2.2. Beta ratio becomes less important as V_{DD} scales down to 0.8V and 0.6V. At these voltages, threshold

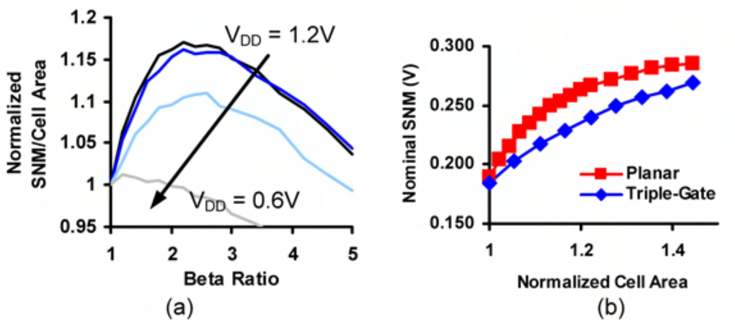

Figure 3.7. Increases in cell beta ratio can improve SNM in the planar cell, but with diminishing returns. Beyond a beta ratio of 2.2, the ratio of SNM to cell area decreases (a). The width of the triple-gate **PD** devices can also be increased, but the beta ratio does not increase proportionally, leading to smaller SNM improvement.

voltage mostly determines the drive strengths of the NMOS transistors. The triple-gate cell can also be manufactured with greater **PD** width, but to limited effect at these dimensions, as illustrated in Fig. 3.7b. For triple-gate devices, increasing the layout width has a proportionally smaller effect on the effective width, which includes a contribution from the device sidewalls. It also increases threshold voltage as the side gates move further from the center of the channel. Although nominal SNM still increases with wider **PD** devices, the gain is less than for the planar cell. The yield benefits are also expected to reduce as the wider **PD** devices begin to resemble more of a planar transistor. Two parallel, minimum-width **PD** devices can be used to increase beta ratio and layout efficiency, but only if $W_{eff} > P$. Cells with very large W_{PD} can use alternative layouts to reduce cell area, if constraints on feature linearity and device orientation are removed. In most cases, though, area efficiency concerns will constrain multi-gate SRAMs to minimum-width devices and unit beta ratios.

That notwithstanding, the multi-gate devices can still provide for superior yields in a smaller cell area. Even if the planar cell is allowed a larger beta ratio, the triple-gate device still has approximately 1σ higher read yield (Fig. 3.8). Although a larger beta ratio increases nominal SNM in the planar cell, it is not enough to compensate for the greater V_{T0} variation and sensitivity to L_G variation. The higher nominal SNM does increase cell sigma, but only up to 1σ.

Following the methodology presented in section 2.4, the triple-gate cell design can be

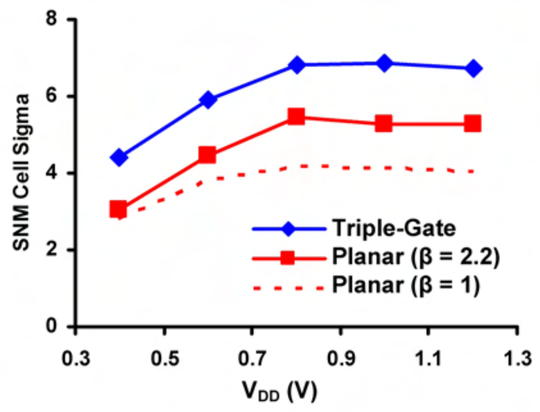

Figure 3.8. Even allowing for a higher beta ratio, the triple-gate cell still has superior yield, due to improved robustness at the device level.

optimized for parameters such as V_{min}. Table 3.2 lists the modifiable device parameters and their allowed ranges. A maximum CD of 20nm was allowed, with CD reductions of up to 30%. A single V_{T0} for all NMOS devices was assumed, which can be tuned by changing the dose or depth of the SSR ground plane profile. A similar allowance is made for PMOS devices. The optimum point for V_{min} was found by evaluating sensitivities to V_{DD} around 0.6V. Effectively, it represents a small shift in favor of a more write-able cell. Fig. 3.9 illustrates read and write cell sigma curves for the optimized design. V_{min} at six sigma yield is reduced to just under 0.6V. In contrast, the planar cell never achieves six sigma yield, at any V_{DD}, and achieves only a four sigma V_{min} at 0.8V.

After yield and cell area, access time is another SRAM metric of importance. Read access time consists of the time needed to develop a high voltage on the wordline at the gate of **PG** and discharge the capacitance on **BL** to a level that can be accurately read out. The latter is the more significant delay, but the bitline capacitance consists mainly of junction and wire capacitance. Since these parasitics are the same in both cells, the read access time correlates directly with the DC current through **PG3** or **PG4** when the low-voltage node is **CH** or **CL**, respectively. Fig. 3.10a illustrates a significant increase in triple-gate read current as a function of V_{DD}.

Like the read access time, the write access time depends strongly on the current through

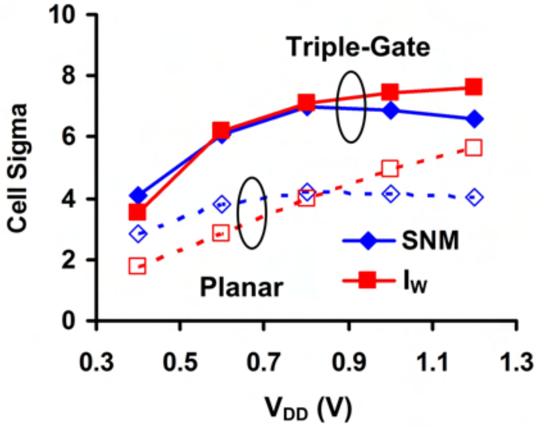

Table 3.2. Allowed Parameter Ranges

Parameter	Allowed Range	Optimum
PD W	14nm - 20nm	20nm
PD L	14nm - 20nm	20nm
PG W	14nm - 20nm	20nm
PG L	14nm - 20nm	20nm
PU W	14nm - 20nm	20nm
PU L	14nm - 20nm	20nm
NMOS V_{T0}	0.31V - 0.37V	0.35V
PMOS V_{T0}	0.34V - 0.40V	0.40V

Figure 3.9. Triple-gate cell design can be optimized for SRAM metrics such as V_{min} by modifying the parameters in Table 3.2. A $V_{min} < 0.6$V at six sigma yield can be achieved in this manner, whereas the planar cell never reaches six sigma yield.

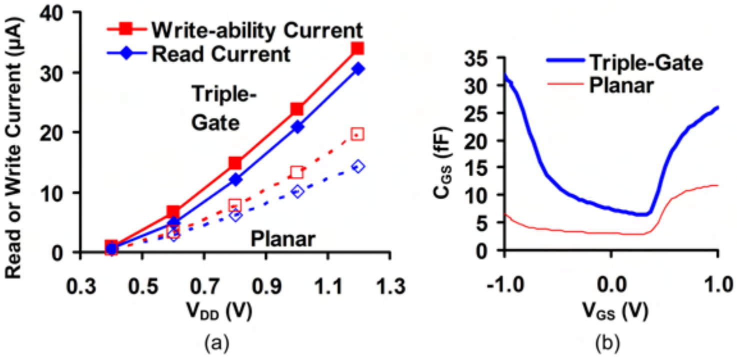

Figure 3.10. Read and write currents are significantly higher in the triple-gate cell (a); however, gate capacitance is also higher (b), offsetting the improvement in write speed. The write-ability current is higher than read current in both cells due to a high floor voltage of $V_{CH} \approx 12\%$ of V_{DD} in the read case.

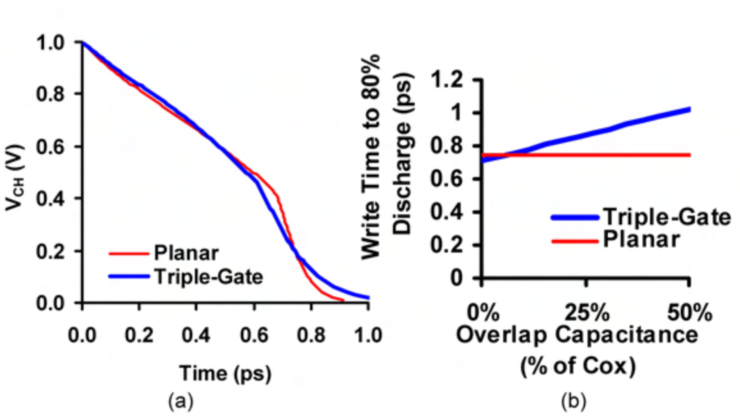

Figure 3.11. Both planar and triple-gate cells discharge at a similar rate (a), even if a moderate additional capacitance is added to the triple-gate's internal node (b).

PG. In writing, though, the capacitance on the discharging node is dominated by gate capacitances, from one of the inverters in the SRAM cell. Although triple-gate devices have an increased drive current, they also have an increased gate capacitance (Fig. 3.10). The gate delay, approximated as $C_{ox}V_{DD}/I_{DSAT}$, is slightly larger than that of the planar device (0.8ps vs. 0.7ps) with equal V_T and $V_{DD} = 1.0$V. This would suggest a slower write access in the triple-gate cell; however, it does not consider the impact of the inverter, which at first resists and later assists the writing of the cell. In order to better approximate the total effect on write-ability, a pseudo-transient simulation was performed (Fig. 3.11). I_{CH} and C-V curves (for NMOS and PMOS) were used to determine the voltage on the internal node **CH** for successive time steps at $V_{DD} = 1.0$V. Initially, the triple-gate cell **CH** discharges a little slower than the planar cell, up to approximately 0.7V. Then, as **PD1** begins turning on, I_{CH} increases much faster in the triple-gate cell, and the discharge rate increases. These changes are slight, though, and the overall discharge time is similar. Actual triple-gate devices are expected to have an extra overlap capacitance between the gate and ground plane to avoid V_{T0} variations due to device height. Fig. 3.11b illustrates the time to discharge **CH** by 80% of V_{DD} as a function of overlap capacitance. Even up to approximately 15% of C_{ox}, the write times are comparable to that of the planar cell. Furthermore, a higher yielding device technology may allow for shorter access time windows for six-sigma cells.

In summary, triple-gate bulk devices are expected to provide a viable device solution to the problem of random dopant fluctuation. By increasing gate control of the channel, triple-gate SRAM cells can exhibit dramatically higher read and write yields, even though nominal SNM is comparable. These benefits come without penalty to access time or cell area; indeed, the triple-gate cell is expected to have faster read access and SNM yield than a larger, optimally sized, planar cell. The design of the triple-gate cell can be optimized to achieve a $V_{min} < 0.6$V at six sigma yield, a full two sigma higher than the planar cell can achieve at any V_{DD}. If the processing challenges–including, for one, the need for tight CD control–can be overcome, then multi-gate devices can significantly extend SRAM scaling.

3.3 Reducing variation from lithography

For sub-20nm devices, lithography variations are expected to be the major source of device variations. Lithography variations affect all of the major device parameters including gate length, channel width, and indirectly threshold voltage. As discussed in Section 2.4, scaled devices are highly sensitive to L_G variations, because of the DIBL effect on V_{TSAT} (Eqn. 2.21). In addition, V_{TLIN} may be sensitive to L_G variations if the device exhibits a steep V_T rolloff behavior, as is common in device processes using large halo implants. Generally, devices have a linear sensitivity to channel width variations ($I_{DSAT} \propto W$), but this can increase with narrow width effects or the adoption of a multi-gate architecture.

Variation in lithography can arise from either the optics or the photoresist. With critical dimensions pushing the resolution limit of optical lithography, printed features can have rounded corners or variable width dependent upon the surrounding features. Modern photomasks include sub-resolution optical proximity correction (OPC) features to reduce these effects significantly, but not entirely. Some pattern degradation remains and can affect device variability in high density patterns, such as those used for SRAM. The variability is of a mostly systematic type, but there is some random variation as well [26]. Recent SRAM cell layouts have evolved to increase linearity in the gate and active layers.

Unlike optically-based variation, roughness variation in the photoresist is mostly random. During the post-exposure bake of a lithography process, the polymers in the resist cross-link and become insoluble. At the edges of the exposed regions, aggregations of cross-linked polymers form a rough edge to the printed features. The size of the aggregations determines the line edge roughness (LER) for each edge independently, with a standard deviation that varies indirectly with feature size. For feature sizes of 100nm, standard deviations as low as $\sigma_{LER} = 3.3$ nm have been achieved [27]. At 20nm, $\sigma_{LER} = 5$ nm has been reported [28]. LER affects CD variation on both edges. The line width roughness (LWR) for a line with uncorrelated edge roughness has a standard deviation of $\sigma_{LWR} = \sigma_{LER}\sqrt{2}$.

3.3.1 Linear Features

The easiest patterns to print are repeating lines and spaces. Line and space patterns have spatial frequency information in one dimension only, perpendicular to the lines, with the majority of the information at the spatial frequency corresponding to the pitch of the pattern. By contrast a checkerboard pattern has information in two dimensions, with the sharpness of the corners defined at high spatial frequencies. A lithography system with a coherent light source behaves like a low-pass filter in spatial frequency, transmitting only the larger, rounder information while filtering out the smaller, sharper parts. Thus a pattern of lines and spaces will be more accurately printed near the center of the lines, while the corners will be rounded.

SRAM bitcell layouts have evolved accordingly (Fig. 3.12). Recent SRAM layouts use long thin cells, with less of a difference in **PD** and **PG** width. An extreme example is the layout of Fig. 3.12b, in which the active and gate regions are almost perfectly straight [29]. Such a cell is expected to have very low variation in device width and gate length, but the tradeoff is a low beta ratio and a low nominal SNM. The cell of Fig. 3.12b uses circuit techniques to recover the lost SNM, which are discussed in Chapter 4. A relatively longer L_{PG} can also be used to maintain the beta ratio and nominal SNM.

Increasing L_{PG} has the two-fold advantage of increasing beta ratio while reducing the

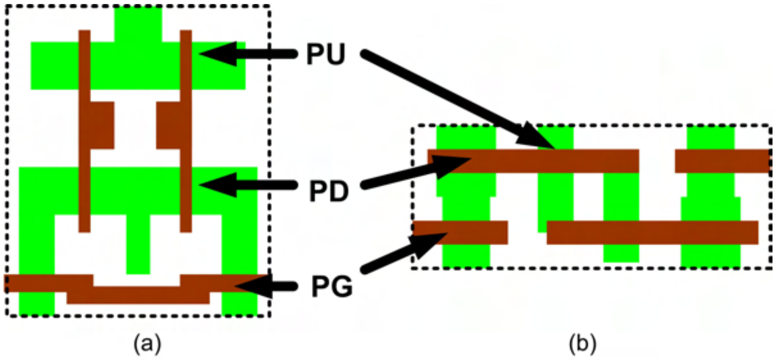

Figure 3.12. Early cell layouts favored a square cell, such as this $2.0\mu\text{m}^2$ cell manufactured in a 130nm technology (a, from [30]). Recently SRAM layouts have become more linear, such as in this $0.346\mu\text{m}^2$ cell in 45nm technology (b, from [29]) with nearly straight active and gate regions to reduce CD variations.

variability. Even a modest increase in L_{PG} can maintain read yield. For example, using the planar cell simulated in section 3.2.2, a cell with $W_{PD} = L_{PD} =$ 20nm devices has approximately 190mV SNM at $V_{DD} =$ 1.0V, 50mV lower than that of a cell with a beta ratio of two. Increasing L_{PG} by 50% to 30nm recovers 30mV of SNM but still leaves a difference of 20mV. However, from eqn. 3.1, an 18% reduction in V_T sigma can be expected from a 50% L_G increase. Simulation of the most probable point of failure reveals that the cell sigma for the design with longer L_G is 6.5, a net yield increase of 0.3σ over the higher beta ratio cell. SRAM cells with straight active layouts can therefore have comparable read yields, even if nominal SNM is lower. They also have smaller area. Of course, the tradeoff is in the write-ability, where in this example cell sigma is 1.1σ lower in the longer L_G case. Whether a straight active design will provide a net benefit to total SRAM yield depends on the particular process and application. For designs where CD variability is the primary concern (e.g. for devices with undoped channels), straight active layouts may allow further yield enhancements from highly uniform patterning technologies such as spacer lithography.

Spacer Lithography

Spacer lithography [31] (also called *sidewall image transfer*) is a patterning technique of recent interest for future technology nodes, because its use of a very uniform and controllable deposition step allows for very thin lines with low CD variation and reduced pitch [32]. In

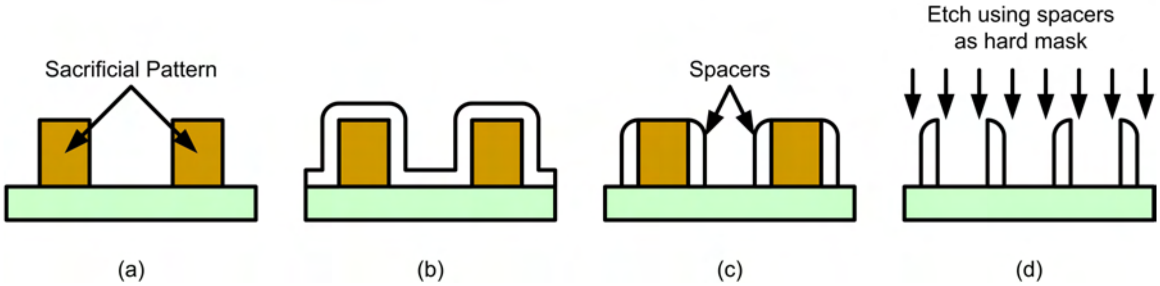

Figure 3.13. A conventional spacer lithography process flow. Following the patterning of a sacrificial material (a), a layer of hard masking material is conformally deposited (b) and anisotropically etched so as to leave only spacers on the sidewalls (c). After the sacrificial material is selectively removed, the spacer-defined pattern can be etched into the substrate (d).

spacer lithography, a chemical vapor deposition (CVD) step is used to form a conformal layer over a sacrificial pattern, usually consisting of repeating lines and spaces (Fig. 3.13). CVD is a very controllable step that results in very uniform thicknesses on top of the sacrificial pattern, as well as along its sidewalls. The substrate is then anisotropically etched so as to remove the CVD layer everywhere, except for spacers along the sidewalls. The sacrificial material is then selectively removed, leaving only the spacer, which are then used as a hard mask to transfer the pattern to the substrate. The LER on each edge of the spacer-defined pattern is correlated, resulting in much lower σ_{LWR} than can be achieved with resist-defined features. Since two spacers are formed at the opposite edges of one sacrificial line, the average pitch from using the spacer process is halved. This enables integration at densities greater than that achievable with photolithography or, alternatively, integration at similar densities using less expensive photomasks with relaxed dimensions.

One complication to the spacer lithography process is that the sidewall spacers formed on opposite edges of the same sacrificial feature are always connected at its ends (Fig. 3.14). An extra lithography step is therefore required to break this connection and trim the spacers to appropriate lengths. In most cases, this extra step does not require critical lithography and can be performed with a second, less expensive mask. However, in the cases where a minimum-sized cut is required, as illustrated in Fig. 3.14a, the maximum integration density that can be achieved can be limited by this step. A second complication

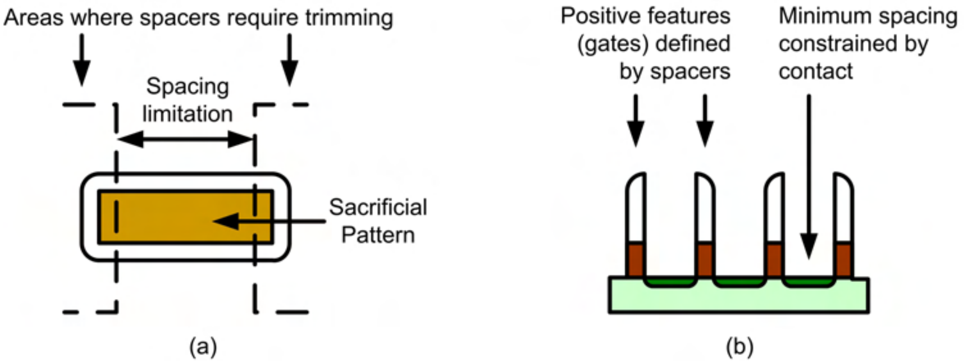
Areas where spacers require trimming
Spacing limitation
Sacrificial Pattern
(a)
Positive features (gates) defined by spacers
Minimum spacing constrained by contact
(b)

Figure 3.14. Spacer lithography processes have required additional conventional photolithography steps to trim the spacers at the edges of the sacrificial pattern (a). This step may constrain integration density if the areas to be trimmed are closer than the minimum spacing allowed by the photolithography. In addition, the minimum pitch for densely-packed transistors may be constrained by the minimum contact pitch if photolithography is required to define these features (b). By contrast, negative spacer lithography enables definition of these regions at smaller pitches than can be achieved with photolithography alone.

is that spacer lithography has been demonstrated until now only for positive features (e.g. lines), where the material under the spacer pattern is protected during an etch. To pattern features such as contact holes, conventional lithography has been required and can limit the minimum achievable pitch (Fig. 3.14b).

3.3.2 Negative Spacer Lithography

The full benefits of using spacer lithography can be realized only if negative features can also be defined with spacers. An extension to spacer processing, called negative spacer lithography, is presented to define these negative features, such as trenches, cut-lines, contact holes, or vias [33]. The negative spacer process is similar to that of conventional, positive spacer lithography, except that the sacrificial material is used as a hard mask. (Fig. 3.15). Following formation of the sidewall spacers, another layer of the original sacrificial material is deposited by CVD and planarized to expose the spacer, e.g by chemical mechanical polishing (CMP) or another means. The spacer is selectively removed to leave a narrow gap. A trench can then be etched through the gap, with the remainder of the substrate protected by the original sacrificial patterns and the second sacrificial deposition. This

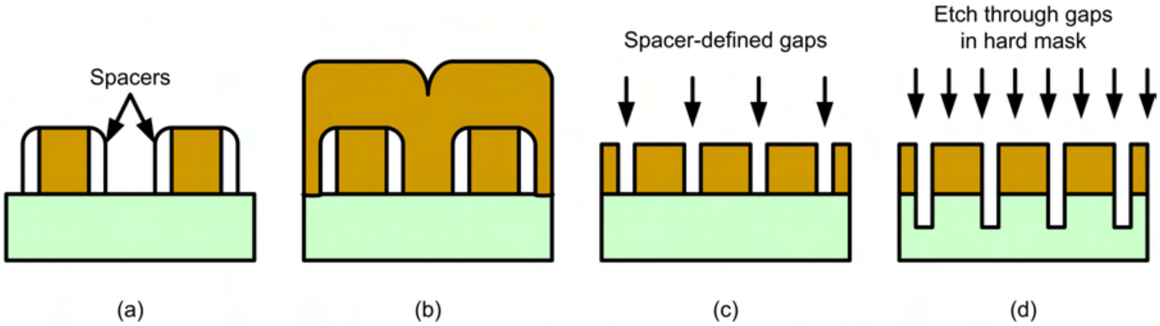

Figure 3.15. A negative spacer process begins by forming spacers around a sacrificial layer (a), as described above. Instead of removing the sacrificial layer, however, a second deposition of the original sacrificial material is performed (b). The substrate is then planarized until the spacers are exposed and can be selectively removed, e.g. with an isotropic etch. The removal of the spacers leaves gaps in the hard mask (c), through which a trench or cut line can be etched.

process is similar to that used for nano-gap capacitors [34], except that the pattern is etched into the substrate. The hard mask may therefore require thinning by a timed etch to avoid aspect-ratio limitations, as described in more detail below.

Trenches and cut lines

Trenches down to 30nm in width are demonstrated on a silicon substrate using negative spacer lithography. First, a 250nm low-temperature-oxide (LTO) etch stop layer is deposited using low pressure (LP) CVD. The sacrificial layer is composed of 400nm thick amorphous silicon, deposited by LPCVD. Amorphous silicon (a-Si) is chosen over polysilicon to reduce the sidewall roughness associated with large grain sizes in a polycrystalline film. The sacrificial layer is patterned with a 0.5μm trench mask, near the lower resolution limit of i-line photolithography. 60nm wide spacers are formed using a 100nm LPCVD recipe for phosphosilicate glass (PSG) which is known to have a step coverage of 60%.

For the second layer of masking material, an etched-back photoresist was investigated first. A thick 1μm layer of i-line photoresist was spun onto the substrate after spacer formation. Using a carefully controlled oxygen plasma, the height of the photoresist was reduced until the top of the a-Si layer and spacers were exposed. Following spacer removal, SEM images of the spacer-defined gaps showed a high degree of roughness along the edge defined by the photoresist. This may be caused by poor adhesion of the photoresist to the sidewall. For low CD variation then, it is desirable to use hard masking material on both edges of the gap. Subsequently in this work, a second deposition of 400nm of amorphous silicon is used.

Although chemical mechanical polishing (CMP) is an effective technique to planarize the substrate and expose the spacers, it is relatively expensive. However, if a line and space pattern of sufficiently high density is used, it is possible to planarize the substrate without CMP and thereby reduce processing cost. First the hard masking material is deposited conformally to such a thickness that the region between the spacers is filled. As a result of the conformal deposition, the material is thicker vertically along the sidewalls, in between the spacers, so that the topography of the region is more planar (Fig. 3.15b). A timed anisotropic etch therefore leaves a small amount of sacrificial material in the planarized region, protecting the substrate between the spacers. The size of the region that can be planarized by this process was empirically found to depend on both the thickness of the spacers and the height of the sacrificial layer, under the following condition:

$$\frac{2}{3}(h-t) > \frac{s}{2} - t \tag{3.5}$$

where h is the height of the initial sacrificial pattern, s is the separation between the spacers (and the width of the region to be planarized), and t is the thickness of the spacers. If h is too small relative to s, no amount of deposition and etch back can planarize the region, since little extra material is deposited on the sidewalls. The maximum h is limited by aspect ratio considerations, discussed in more detail below. If a suitable sacrificial layer height and pattern spacing can be used, planarization in this manner can be considerably cheaper than CMP. Otherwise, an alternative technique such as CMP is necessary to planarize the substrate.

Following exposure of the spacers using this planarization process, the spacers are selectively removed with a dilute HF dip. The 100:1 dilution of HF has a selectivity of PSG to undoped LTO of approximately 5:1, allowing for the removal of the spacers with minimal damage to the underlying oxide layer. An overetch of 100% is used in the spacer removal step to ensure complete evacuation of the gaps. In Fig. 3.16a, a top-down SEM image shows two spacer-defined gaps in the amorphous silicon hard mask. The path of the gaps is curvy because the initial sacrificial edge was rough; however, the width of the gaps remains constant. Small cavities in the planarized hard mask between the spacer-defined features are seen where the gaps were too far apart for planarization following eqn. 3.5 above.

The oxide layer is then etched anisotropically through the gaps using a CHF_3 plasma at 200mTorr and 700W. This highly selective and anisotropic etch transfers the negative pattern of the gaps into the oxide layer immediately above the substrate. The silicon substrate is etched with a Cl_2 and HBr plasma, which also removes the remaining sacrificial layer. Finally, the oxide layer is selectively removed in a solution of dilute HF to leave only the silicon substrate and the etched trenches.

Fig. 3.16b shows a cross-sectional SEM image of the resulting trenches, with depth of 860nm and width of 60nm. Other samples processed with a thinner spacer layer result in trenches of 30nm width and 490nm depth for the same silicon etch recipe (Fig. 3.16c). The difference in etch rates and thinning at the bottom of the trenches can be attributed to aspect ratio dependent etching (ARDE) and micro-loading effects.

The decrease in etch rate is one of the aforementioned aspect ratio considerations that limits the maximum sacrificial layer thickness. A thick hard mask with thin, high aspect ratio gaps will have a slower etch rate at the substrate. This will effectively reduce the etch selectivity to the sacrificial layer or, alternatively, reduce the maximum trench depth. To alleviate this issue, the hard mask material can be removed prior to the substrate etch. The underlying oxide etch stop layer is then used as a hard mask.

A second aspect ratio consideration is the control of the gap width (Fig. 3.17). If the sacrificial layer and spacer etches are not perfectly anisotropic, the final trench may be narrowed by shadowing. From purely geometrical reasoning, the trench width in such a condition is expected to follow:

$$w_{gap} = t - \frac{h}{\tan \Phi} \tag{3.6}$$

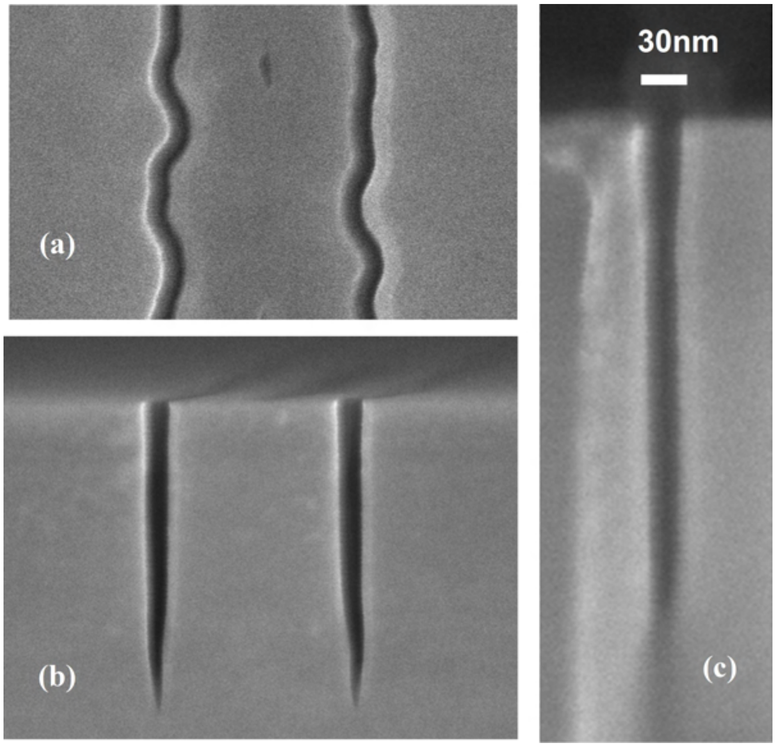

Figure 3.16. SEMs during negative spacer lithography processes: after spacer removal (a, top down) and after substrate etch (b & c, cross-sectional). Even though the lithography edge is rough, the spacer-defined trench has excellent CD uniformity (a). With the appropriate sacrificial layer height, the area between the spacers can be planarized via CVD and reactive ion etching (RIE), avoiding the need for CMP (b). Negative features down to 30nm have been fabricated with this process (c).

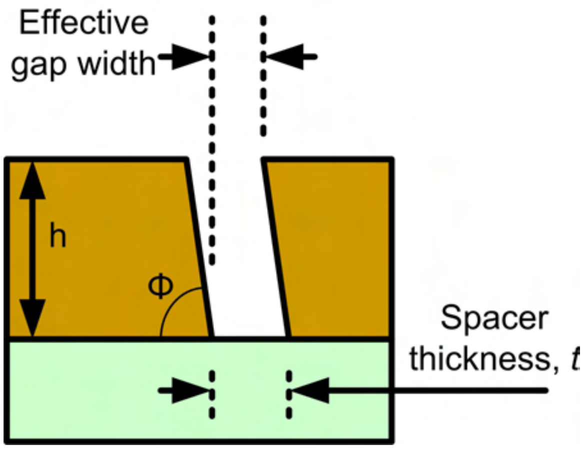

Figure 3.17. If the etch process is not perfectly anisotropic, sloped sidewalls on the sacrificial layer can cause sloped spacers, which in turn can narrow the effective gap width. This presents an additional source of variation for features defined with negative spacer lithography, but it can be mitigated by reducing the height of the sacrificial layer.

where w_{gap} is the effective gap width, and Φ is the angle of the sidewall. With a high aspect ratio gap ($h \gg t$), greater control of Φ is required in order to reduce variation in the trench width. Thinner hard masks therefore reduce width variation as well as enable deeper trenches.

Contact holes and vias

Contact holes can be made with two overlaid negative spacer processes (Fig. 3.18). The process begins as above to make a planarized hard mask with the spacers exposed. The spacers (Spacer1) are not immediately removed, however. Instead, another layer of sacrificial material is deposited and patterned. Then a second set of spacers (Spacer2) is formed. Another hard mask layer is deposited and planarized to expose the spacers. Using an isotropic etch, e.g. dilute HF for PSG spacers, the spacers from both layers are selectively removed to leave gaps. The substrate is exposed only at the intersections of each gap. An anisotropic etch through the gap will create a hole in the substrate.

Fig. 3.19 illustrates this process with top down SEM images. A substrate of 100nm LTO on silicon is used with a 200nm polysilicon sacrificial layer as the hard mask. For improved visibility the hard masks are left unplanarized. The spacers on both layers are formed of 60nm PSG. Fig. 3.19b shows the process after the spacers are exposed by anisotropic etch back, leaving a second, outer spacer of polysilicon outside the 60nm PSG Spacer2. The underlying layer can be seen with its exposed spacer (Spacer1) on the left of the image. The spacers are removed with dilute HF, and the underlying oxide is etched with a CF4 plasma. Fig. 3.19c shows the region circled in Fig. 3.19b after the sacrificial polysilicon layers are removed. A 60nm hole is visible at the intersection of the spacers. Where the underlying Spacer1 was exposed, a trench has been cut into the oxide.

The same process can be used to make a dense grid of holes by using sacrificial patterns with dense lines and spaces. The resulting holes are expected to have the same pitch and uniformity improvements as lines defined by conventional spacer lithography. These improvements are important to high density device integration in two ways. First, the

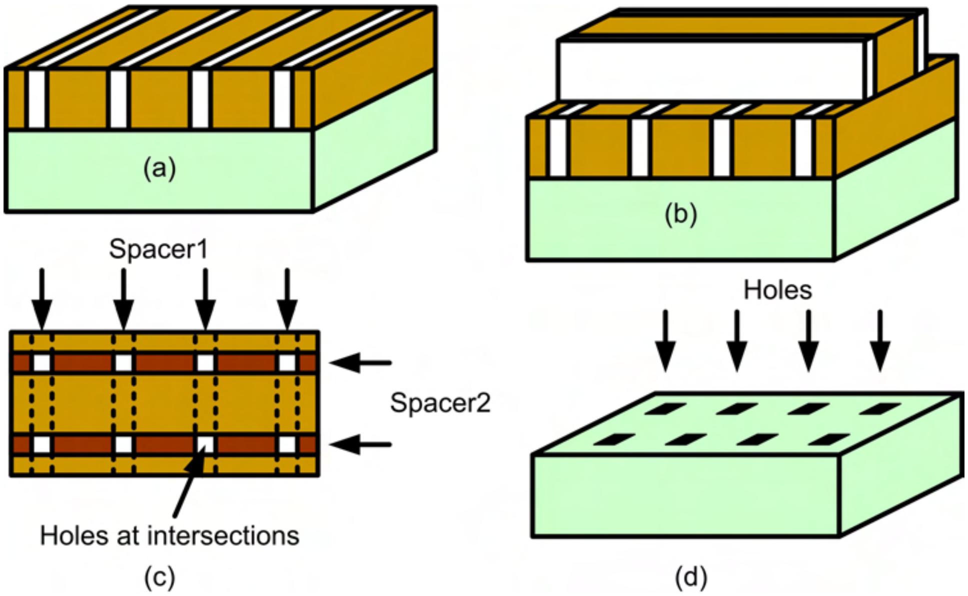

Figure 3.18. A negative spacer lithography process for producing contact holes and vias uses two sets of spacers. After exposing the first set (Spacer1) in a planarized sacrificial layer (a), a second hard mask with intersecting spacers (Spacer2) is created (b). After a deposition of sacrificial material and planarization, the spacers are selectively removed, exposing the substrate only at the intersections of Spacer1 and Spacer2 (c, top-down), through which contact holes or vias can be etched (d).

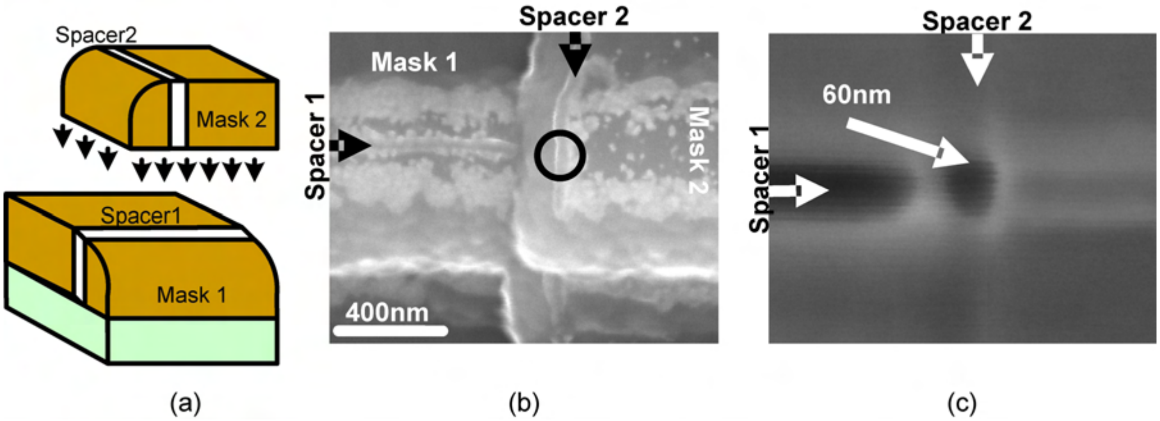

Figure 3.19. Contact holes created with the process described in Fig. 6 using two orthogonal 60nm trench masks and no planarization, as illustrated (a) and seen via top-down SEM (b). The spacers are selectively removed, leaving a 60nm hole at the intersection (c).

full benefits of pitch reduction can be realized, allowing integration at four times greater density than that achievable with photolithography. Second, uniformity improvements in the size of contact holes reduce variations in the contact resistance. Contact resistance can affect transistor V_{TSAT} as well as other device parameters. Variability in the contact resistance can cause asymmetric source/drain behavior, a phenomenon to which SRAM cells are particularly sensitive in the pass-gate transistors. The costs of negative spacer lithography are the additional lithography, CVD, and etch processing required to pattern negative features, and possibly added mask costs. (However, the discussion in Section 3.4 will illustrate ways in which many masks can be reused for several steps.) Alignment of the sacrifical patterns remains a growing challenge as tighter pitches are used, but no more so than that for photolithography. Although two alignments are necessary for making holes with a negative spacer lithography process, each alignment is critical in only one direction. The overall alignment tolerances are therefore expected to be similar. Although the processing costs can be considerable, negative spacer lithography may prove to be a less expensive (and less variable) alternative to e-beam lithography or extreme ultraviolet (EUV) systems for continued scaling.

3.3.3 Iterative Spacer Lithography

Spacer lithography can be iterated to achieve ultra high densities. A single spacer step reduces the average pitch in a line and space pattern by a factor of two. With n iterations, the reduction is 2^n. To achieve a final pattern of width W_n with equal spacing S_n between all lines, the width of the sacrificial pattern in the preceding step must have a width $W_{n-1} = S_n$ and spacing of $S_{n-1} = S_n + 2W_n$. This can result in initial patterns with a duty cycle significantly different from that of the final pattern. The relationship for several iterations is presented for quick reference in Table 3.3.

Iterated processes have been demonstrated by forming spacers on the sidewalls of existing spacers [34, 35]. This approach may introduce additional variation in line widths, if the etch processing is not perfectly anisotropic. As for spacer-defined gaps, the final

Table 3.3. Dimensions of Sacrificial Features Necessary to Create Final Pattern of W_n, S_n

No. Steps	Initial Width	Initial Spacing	Initial Pitch
n	W_0	S_0	$P_0 = W_0 + S_0$
4	$5S_n + 7W_n$	$11S_n + 10W_n$	$16P_n$
3	$3S_n + 2W_n$	$5S_n + 7W_n$	$8P_n$
2	$S_n + 2W_n$	$3S_n + 2W_n$	$4P_n$
1	S_n	$S_n + 2W_n$	$2P_n$

thickness of the spacer-defined feature is determined not only by the thickness of the spacer t, but also to some extent by its height h_{sp} and the sidewall angle Φ, similar to eqn. 3.6:

$$W_{feature} = t + \frac{h}{\tan\Phi} \tag{3.7}$$

The additional term can be significant. Spacer heights are usually in the range of $1.5t < h_{sp} < 4t$. The lower limit is set higher than t to avoid variation in feature width due to rounding. (The top of a spacer feature can be described by a quarter-arc with radius t.) The upper limit is determined by the maximum stable aspect ratio. Spacers that are too tall can be physically unwieldy and prone to collapse. The sidewall angle Φ in an anisotropic etch is usually large (near 90 deg), but can vary several degrees. A spacer-defined feature with $h_{sp} = 2W_{sp}$ and $\Phi = 83$ deg will therefore be 25% larger than that expected from W_{sp} alone. Fortunately, this variation will tend to affect all spacers systematically. The patterned sacrificial layer will tend to be symmetric, so spacers on opposite sidewalls will have the same h and Φ. The process can be adjusted to compensate for the width increase, for example by decreasing t.

With a second iteration, though, the sacrificial feature may look symmetric. If the sacrificial layer is simply the spacer of the previous step, it will have a rounded top. The height of one side of the spacer is greater than the height of the other, by approximately t_1. The second iteration of spacers will therefore have different heights, resulting in a systematic variation. For $\Phi < 90$ degrees, the spacer on the taller side will result in a wider pattern.

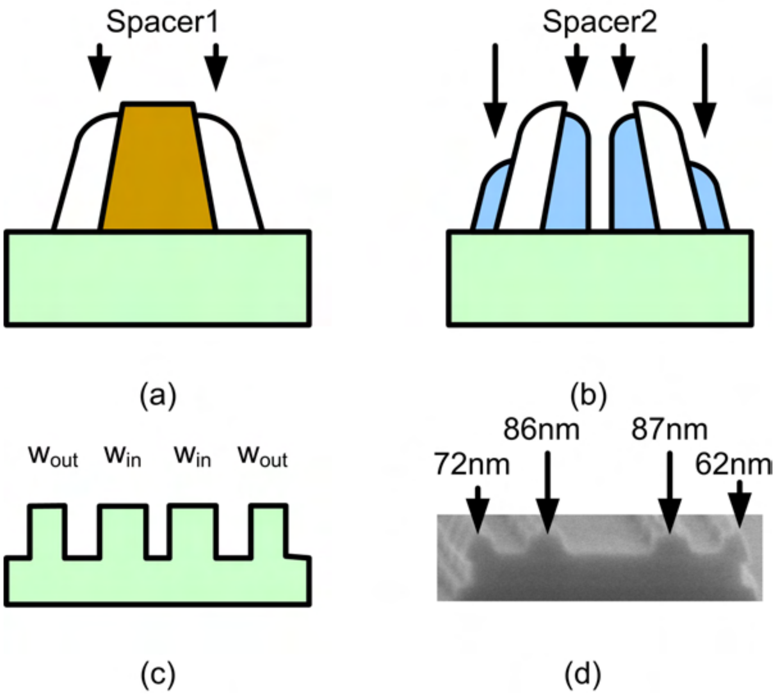

Figure 3.20. An iterative spacer process that forms a second set of spacers on the sidewalls of the first is subject to a systematic offset in final CD, dependent on the thickness and angle of Spacer1 (a-c). The effect was observed experimentally (d).

From Eqn. 3.7, the mismatch will be:

$$\Delta W_{feature} = \frac{\Delta h_{sp}}{\tan \Phi} \tag{3.8}$$

This effect is illustrated in Fig. 3.20. Lines of 100nm width were etched into a 200nm thick polysilicon layer. Although the etch was performed using a nominally anisotropic recipe of Cl_2 and HBr, the actual sidewall angle was measured at $\Phi = 79\,\text{deg}$, possibly due to sloping in the photoresist mask. A $W_{sp1} = 100$ nm phosphosilicate glass (PSG) spacer was formed on the sidewalls, and the sacrificial layer was removed with a highly selective anisotropic polysilicon etch. A second spacer of polysilicon and $W_{sp2} = 54$ nm width was formed on the sidewalls, and the original PSG spacer was isotropically removed in an HF solution. The final features were then etched into a layer of polysilicon below. From equation 3.7, the predicted feature widths are 73 nm and 93 nm on the outside and inside of the pattern, respectively. Using cross-sectional SEM, the measured feature widths

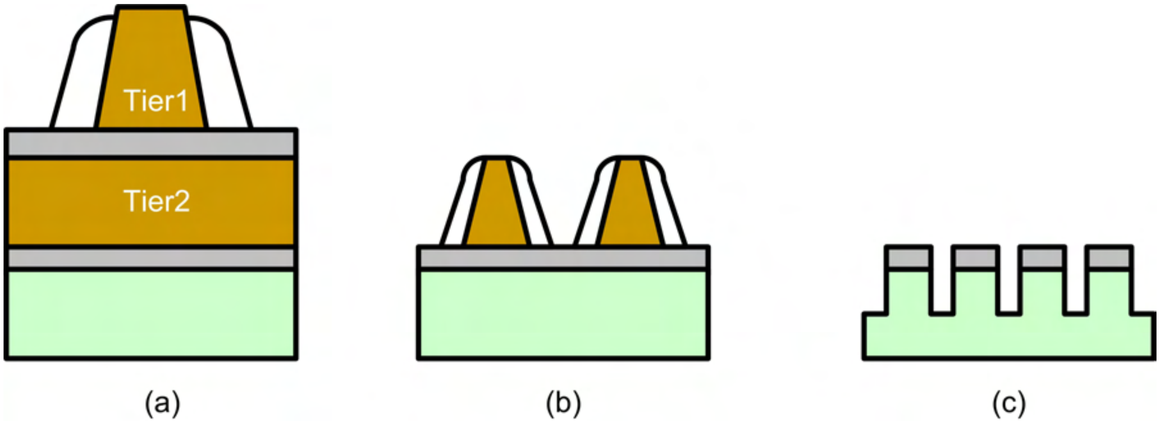

Figure 3.21. An iterative spacer process using a multi-tiered hard mask can remove the systematic offset in CD by ensuring that each set of spacers is formed along a sidewall of equal slope and height. Note that a sloped sidewall etch can still introduce error in the CD of the final features, but the offset is eliminated. Eliminating the offset is important for patterning circuits with high sensitivities to mismatches, such as SRAM bitcells and sense amplifiers.

were actually 67 nm and 86 nm on average. The difference of $\Delta W_{feature} = 19$ nm between the inside and outside features matches well with the 19 nm estimation from equation 3.8.

To mitigate this systematic variation, a multi-tiered hard mask can be used (Fig. 3.21). Each tier of the hard mask consists of a sacrificial layer and a thin etch-stop layer. The use of an etch-stop layer ensures that the sidewall profile of the sacrificial layer does not change with subsequent iterations. Only one set of spacers is formed per tier; the pattern is etched into the next tier and the initial spacers are removed before the next set of spacers are formed. This ensures that every spacer is formed along sidewalls of identical heights and angles. A sloped sidewall etch can still introduce a systematic error in the average feature width; however, the mismatch component of the error is eliminated. This is particularly important for patterning circuits with high sensitivities to mismatch variations, such as SRAM bitcells or differential sense amplifiers. It enables key dimensions of paired devices to be defined by spacers formed during the same step without bias in the final CD.

A three iteration process is illustrated by SEM in Fig. 3.22. An initial pattern of W $= 0.5\mu$m and P $= 1.2\mu$m was defined in a 400nm amorphous silicon layer (Fig. 3.22a). 100nm LTO spacers were formed, the first sacrifical tier was selectively removed, and the

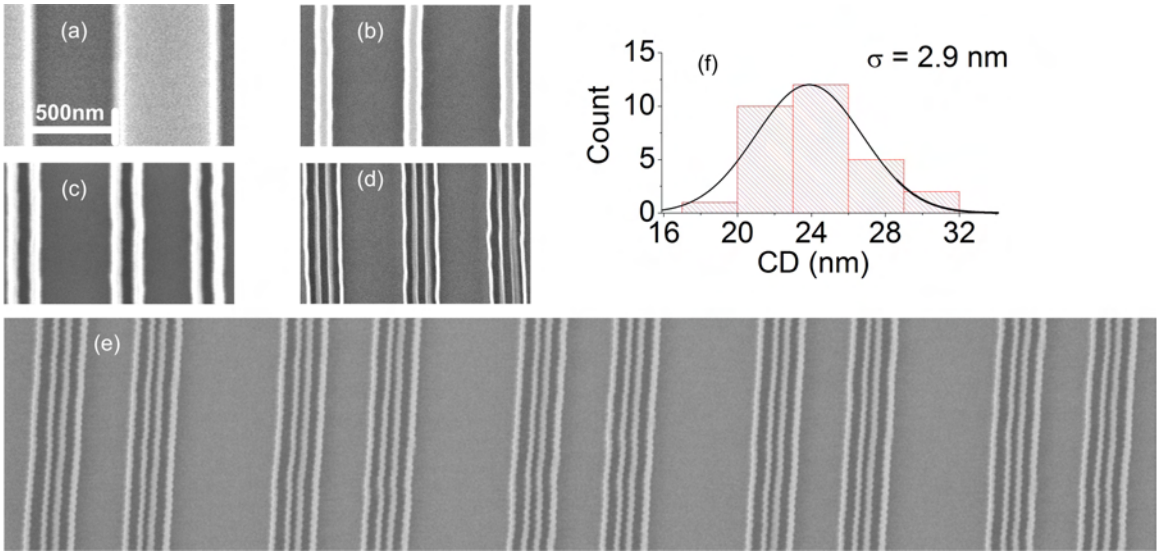

Figure 3.22. A refined spacer lithography process has been developed to allow multiple iterations down to line widths of 20nm with 60nm spacing. From an initial line and space pattern of 1.2m pitch defined with optical lithography (a), successive spacer steps of 100nm (b), 60nm (c), and 20nm (d) are performed to reach the target pattern. Using multiple layered hard masks enables excellent uniformity across the wafer (e). Very low CD variation of $\sigma_{LWR} = 2.9$nm was observed (f).

pattern was etched into the second tier below. 60nm spacers were formed around the 150nm sacrificial layer of the second tier. On the third tier, 20nm spacers were formed around an 80nm thick sacrificial layer. The final 20nm line pattern with 80nm local pitch was etched into oxide to improve the imaging contrast with silicon. Very low line width variation (σ_{CD} = 2.9nm) was measured across the wafer.

Iterated spacer lithography using a multi-tiered hard mask is an effective way to print features at resolutions far below the photolithographic limit. It can be used to define line widths down to 20nm and possibly thinner with less CD variation than that of resist-based masks. As for negative spacer lithography, the cost of an iterated process is mostly in the additional processing required, a relatively inexpensive premium compared to the cost of a 32 nm photomask set ($6 million, projected from recent trends [36]). A summary of the required processing is presented in Table 3.4, compared with an alternative double patterning process using multiple photolithography steps. The table summarizes the pitch and variation benefits as well as the required processing needed for one dimensional features, Table 3.4. 1-D Double Patterning and Spacer Lithography Process Comparison (Expensive steps highlighted in **bold**)

Process	Double Patterning	Negative Spacer	Iterated Spacer
Pitch	1/2	1/2	1/4
σ_{CD}	$\sigma_{LER}\sqrt{2}$	σ_{LER}	σ_{LER}
			Hard Mask Tier 1 CVD
			Etch Stop CVD
	Hard Mask CVD	Hard Mask CVD 1	Hard Mask Tier 2 CVD
	Photolithography	**Photolithography**	**Photolithography**
	Trim	Spacer CVD 1	Spacer CVD 1
	Etch	Hard Mask CVD 2	Etch
	Photolithography	**Planarization**	
	Trim		Spacer CVD 2
	Etch	Etch	Etch

such as lines and spaces. The $1/\sqrt{2}$ reduction in CD variation is expected due to the correlated line edge roughness of a spacer process; however, it comes at a processing cost. To define the second dimension, the negative and iterated spacer processes will require repetition of these steps. The double patterning process will need to be repeated to achieve the same 1/4 pitch reduction as iterated spacer lithography. In addition, there can be design costs associated with iterated spacer lithography, since layouts must be composed of patterns that can be repeated several times. This imposes a practical constraint of linearity and regularity on circuit layouts, which may increase design time and cost. Nevertheless, for highly regular and linear circuits, such as SRAMs, the combination of iterated and negative spacer lithography processes is a promising approach for continued SRAM scaling.

3.4 SRAM Design

Two sources of variation, random dopant fluctuation and CD variation, have been identified as especially problematic for SRAM scaling due to their effects on device behavior. Each of these sources can be addressed directly by changes in device or processing technology. Multi-gate devices such as FinFETs or triple-gate FETs will enable undoped or very lightly doped channels, with significant reductions in V_T variation and corresponding improvements to SRAM reliability. Spacer lithography will reduce variability in active, gate, and contact dimensions, allowing further improvements in yield. The two approaches can be implemented simultaneously; indeed, the heightened sensitivity of multi-gate devices to active width variations improves the effectiveness of spacer lithography.

Each approach has a challenge associated with SRAM cell layout. As discussed in section 3.2.2, it is inefficient to achieve a high beta ratio with multi-gate devices due to the significant conduction along the device sidewalls. Beta ratios in such cells can be achieved with longer gate lengths or higher V_T in the **PG** devices, at the cost of write-ability. Spacer lithography, especially when iterated, is best-suited for cell layouts with regular, repeating, and linear features. As described above, modern SRAM layouts already rely on mostly linear active and gate patterns to reduce variations; however, with spacer-defined active or gate layers, the cell beta ratio cannot be adjusted by device sizing. Nevertheless, as discussed in section 3.3.1, SRAM yield may still be improved if the reduction in parameter variation is greater than the decrease in nominal SNM.

A simple means of implementing both approaches is illustrated in Fig. 3.23. Spacer lithography is used to define four parallel strips of active region. Since the same CVD step defines all of the device widths, mismatch variations are minimized, even if the edges of the sacrificial pattern are rough. The spacers are trimmed to form only the six transistors in the bitcell. A conventional lithography step can then be used to produce landing pads for contacts, to improve manufacturability and reduce contact resistance. The gates of the transistors can also be defined with spacer lithography. FinFET SRAMs were first implemented with spacers in [35], in which spacer-defined multi-fin devices were used for all transistors and selective epitaxial growth was used to increase contact area. Although increased SNM variation was reported for the spacer-defined fins, the process suffered from a poor gate stack deposition (another processing challenge associated with three-dimensional

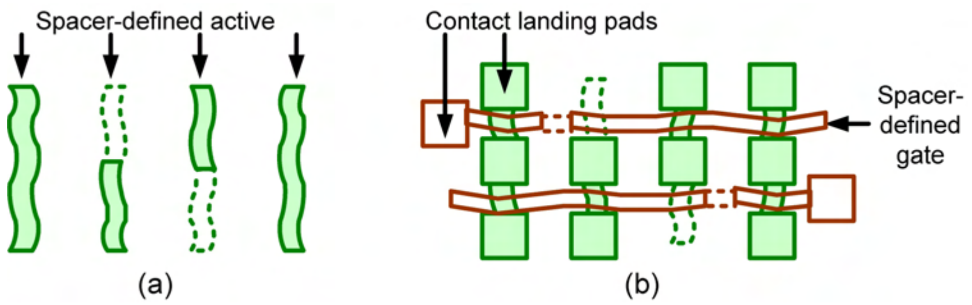

Figure 3.23. SRAMs with spacer-defined device dimensions are expected to have low device variability due to CD variation, even if the edges of the sacrificial layer are rough. In one implementation of a spacer SRAM process (a), the active layer can be defined from spacers off two parallel lines and trimmed to create only six transistors per bitcell (dotted lines). The gate layer can be defined in the same way (b). To ease manufacturing, additional lithography steps can be used to define contact landing pads.

device structures). The resulting FinFETs exhibited a severely skewed V_T and do not accurately represent the yield benefits of spacer-defined fins.

If the multi-gate process is well-controlled and SRAM variation is not the limiting factor for scaling, the minimum cell size will be limited by the minimum pitch of the contact lithography. If negative spacer lithography is used to pattern the contacts, the minimum cell size may still be limited by the conventional lithography needed to pattern the metal layers and local interconnects. The pitch limitations can be completely removed only if every pattern in the cell is defined by a spacer process.

3.4.1 All-Spacer FinFET SRAM

An all-spacer layout is presented in Figs. 3.24-3.28. Every pattern can be defined by conventional or negative spacer lithography from linear features, allowing line pitches below 100nm with iteration. The extreme regularity and linearity of the layout could also minimize variation for a conventional lithography process. The layout is presented specifically for undoped SOI FinFET devices, which benefit cumulatively from reduced σ_{VT0} and CD variations. However, this layout in whole or in part could also be used with other device technologies, including planar MOSFETs.

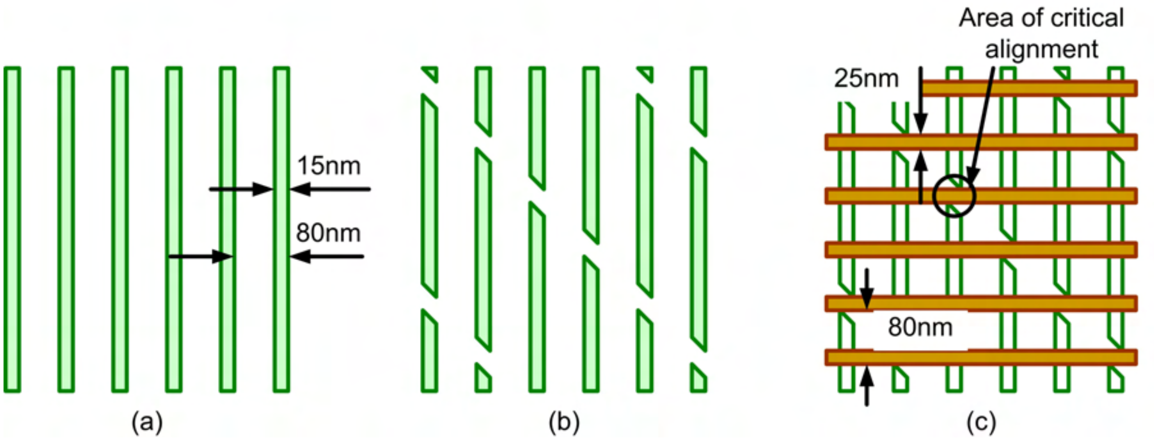

Figure 3.24. The initial active and gate layers of a FinFET SRAM can be patterned with spacer lithography only in the following manner. Vertical lines of active material are patterned at a pitch P (80nm as illustrated in (a)). The active region is cut diagonally (b). After gate stack formation, the gate is patterned into horizontal lines of equal pitch (c). The gate pattern must be aligned to the active pattern as shown.

Initially, regularly spaced parallel lines are defined in a hard mask layer of high temperature LPCVD SiO_2 at 15-20nm width and $P = 80$nm pitch (Fig. 3.24a). Although the final active width specification is 15nm, the width can shrink a couple of nanometers during the etch into the silicon or during gate oxidation. The hard mask pattern is then cut using a negative spacer process at an angle of -45 degrees (Fig. 3.24b). In addition to disconnecting the spacers at the edges of the array, the cut also removes active material from the locations of the inverter gate contacts. The width of the cut should therefore be wide enough that the gate can pass through the gap, at least $L_G\sqrt{2}$ or 35nm for $L_G = 25$nm patterned gate width. The pattern is then etched into the SOI layer, and the gate stack is formed.

The gate layer is patterned with spacers at an orthogonal direction to the active layer, with 25nm width and 80nm pitch (Fig. 3.24c). The gate layer is the first layer with critical alignment. Misalignment of a few nanometers may result in a partial overlap of the inverter gates with the edges of the active region in the indicated area. In the final SRAM cell this could have the effect of increasing the capacitance between the internal node and the bitline of the neighboring column. In the absence of processing defects, though, a misalignment

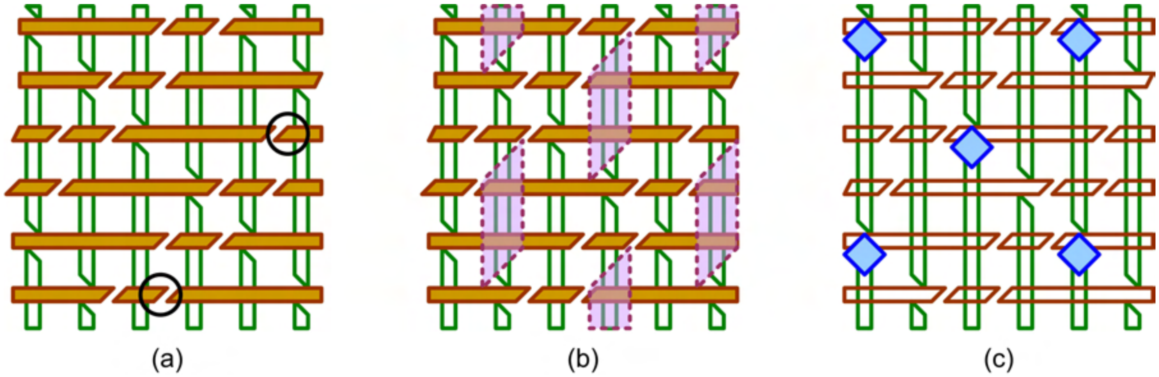

Figure 3.25. Definition of an all-spacer SRAM continues with separation of the wordline gates from the inverter gates (a). Areas of critical horizontal alignment are circled. Source and drain formation can be accomplished with a hard mask and a pattern separating PMOS from NMOS transistors (b). Local interconnects are made to connect the gate of one inverter to the output of the opposite inverter.

of up to 30nm can be tolerated before causing a hard failure. The gate layer is then cut diagonally at an angle of 45 degrees using a negative spacer process (Fig. 3.25a). This cut separates the wordline from the gate of the inverter. The cut need only be wide enough to disconnect the two gates. A 15nm cut is illustrated on both sides of the **PG** transistors. Alignment errors of approximately +/- 10nm can be tolerated without affecting the **PG** transistor. Following gate patterning and sidewall spacer formation, the source and drain implants are performed. An implantation mask can also be generated with a negative spacer process, as in Fig. 3.25b, where holes for the PMOS implant are indicated. The pattern is generated with a $P\sqrt{2} \approx 110$nm wide, $2P\sqrt{2} \approx 220$nm pitch line at 45 degrees and a 60nm vertical line at 160nm pitch. NMOS source and drains can be formed using the inverse pattern as a mask or via counterdoping.

An advantage of using SOI multi-gate or planar devices with this layout is that it is not necessary to include well contacts within the cells. In some SRAM layouts, the wells form continuous strips throughout the array and can be contacted on the periphery. Such layouts may also be partly constrained by minimum n-p device spacing. In this example, for SOI FinFETs, no wells are required. The PMOS implant layer is broken up into patches so that **VDD** and **GND** can be routed linearly.

Following the S/D formation and activation annealing steps, an isolation layer is

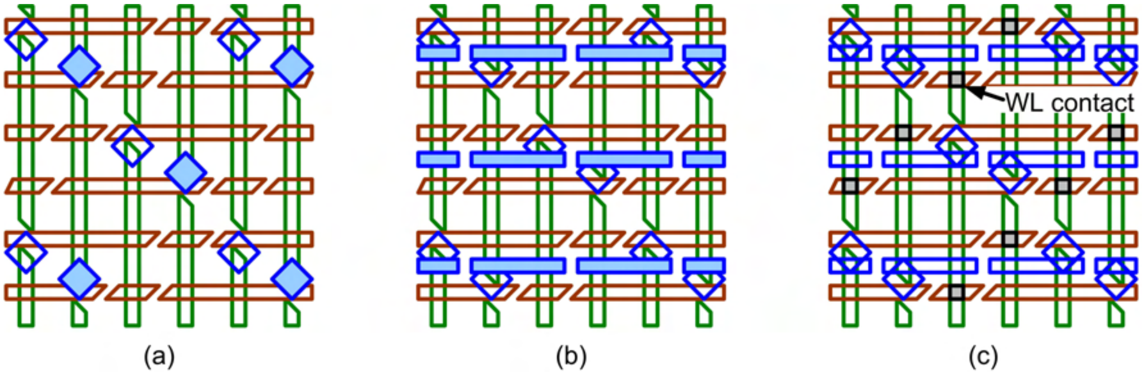

Figure 3.26. A second local interconnect pattern is required to cross-couple the inverters (a) and the drains must be connected (b). Following metal fill of the interconnect patterns and a deposition of an isolation layer, the wordline contact must be patterned to connect to the top of the **PG** transistors (c). This is the most challenging step in the all-spacer SRAM process.

deposited. Holes are then opened for local interconnect formation at the drains of the inverter. In Fig. 3.25c, one set of 45nm holes is opened to connect the gate of one inverter to the drain of the opposite inverter. The holes can be patterned with 45nm negative spacer-defined lines at $2P\sqrt{2} \approx 220$nm pitch and angles of +45 and -45 degrees. Two such holes are needed in the SRAM cell, one for each inverter. Due to the constraints of linearity and regularity of spacer-defined features, the two holes cannot be made simultaneously without distorting their shape. A second set of holes is made with an identical pattern and process to the first, but an offset of (80nm, -40nm) onto the second inverter (Fig. 3.26a). Finally a thin hole is opened in the horizontal direction to bridge the drains of each inverter (Fig. 3.26b). This hole can be patterned with a combination of a horizontal, negative spacer pattern of less than 30nm width and $2P = 160$nm pitch and a vertical positive spacer pattern of less than 40nm width and similar pitch, as shown. If these three local interconnect patterns are etched into a thin hard mask layer over the isolation dielectric, they can be etched as a single set of contact holes with minimal damage to the underlying buried oxide layer or the sidewall spacers around the gate. A metallization process can then be used to fill the holes.

The most challenging step of the all-spacer SRAM process is the formation of the wordline contacts (Fig. 3.26c). Because of the constraint of linearity, it is necessary to

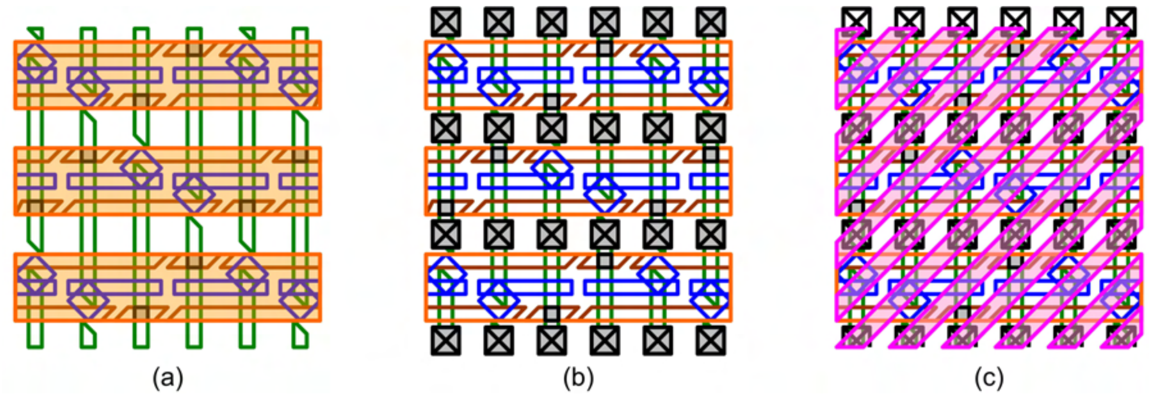

Figure 3.27. The wordline is routed horizontally (a). Contacts are then formed at the top and bottom of each cell (b) to connect to the supply and bitlines, which are routed at an angle of 45 degrees (c).

make a very short gate pattern for the **PG** devices. Unlike in other SRAM designs where the wordline contact is made on the edge of the cell next to the device, in this design it is necessary to contact the gate over the transistor itself. As a result the contact must be very small (up to $L_G = 25$nm) and precisely aligned. Fortunately, as a gate contact, it is less sensitive to variations in resistance than other contacts (e.g. **BL**). Once the contacts are made, the wordline can be patterned with a relatively large horizontal pattern of 100nm width and $2P = 160$nm pitch (Fig. 3.27a), leaving 60nm for the external contacts. The external contacts are regularly spaced contact holes at the edges of the SRAM cell which make connections to **VDD**, **BL**, **BLB**, and **GND** (Fig. 3.27b). Forty nanometer wide contact holes are illustrated, but the actual dimensions can be varied such that the sum of the contact width and the vertical alignment tolerance is less than 60nm. Metal wires are then patterned at an angle of 45 degrees and $P/\sqrt{2} \approx 60$nm pitch to form diagonal columns in the array (Fig. 3.27c).

The final SRAM bitcell and array configuration is illustrated in Fig. 3.28. The bitlines are interleaved with those of the neighboring cells in the same row, such that **BLB** for the neighboring column on the left runs between **BL** and **VDD** and **BL** for the neighboring column on the right runs between **GND** and **BLB**. Each bitcell is rectangular, but since the columns run diagonally, the final shape of the array is a parallelogram. This is illustrated

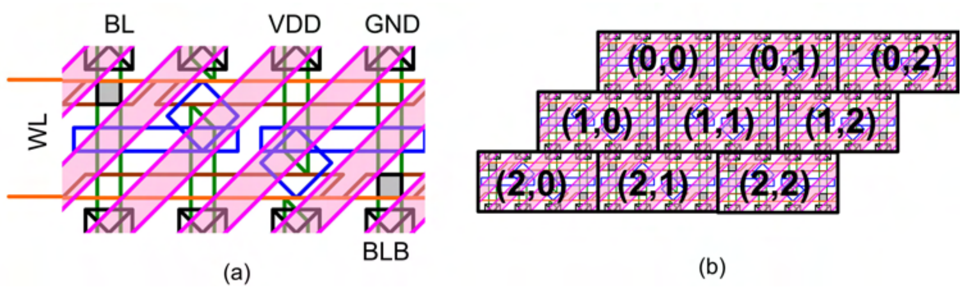

Figure 3.28. The final 0.0512μm^2, 82F^2 SRAM bitcell is rectangular, with bitlines interleaved with those of the neighboring columns (a). Rows run horizontally and columns run at 45 degrees, resulting in a parallelogram-shaped array, numbered as (row, column) (b).

for a 3 × 3 configuration of cells in Fig. 3.28b with row and column numbering. The area of each bitcell is $8P^2 \approx 0.0512\mu\text{m}^2$, or approximately 82F^2 (using $F = L_G = 25$nm).

The cell can be so small because two major obstacles to scaling have been removed. First, the density improvements of an all-spacer process allow patterning at sub-lithographic pitch. Secondly, the use of spacer-defined undoped FinFETs allow for larger yields with a small cell. The major challenge to this process is integrating it with other circuits on the chip. Separate lithography steps are required for the sacrificial patterns in the SRAM array and the circuits elsewhere on the chip. The many extra steps associated with this process make the all-spacer SRAM an expensive, though viable, solution for continued SRAM scaling.

3.4.2 Other circuit layouts

The costs of an all-spacer process can be reduced if an overlap can be found between SRAM and logic patterning steps. Most logic circuits are less sensitive to variation because they use larger devices and have wider noise margins. Nevertheless, spacer lithography can also be used for the critical layers of other circuits, such as standard logic cells.

An example of such a layout is shown in Fig. 3.29, in which the critical dimensions of the active, gate, and contact layers can all be defined exclusively with spacer lithography. The active layer can be defined with horizontal spacers off of an ashed or trimmed pattern from Mask I. Strips of spacer-defined active regions are then isolated with negative spacer-defined

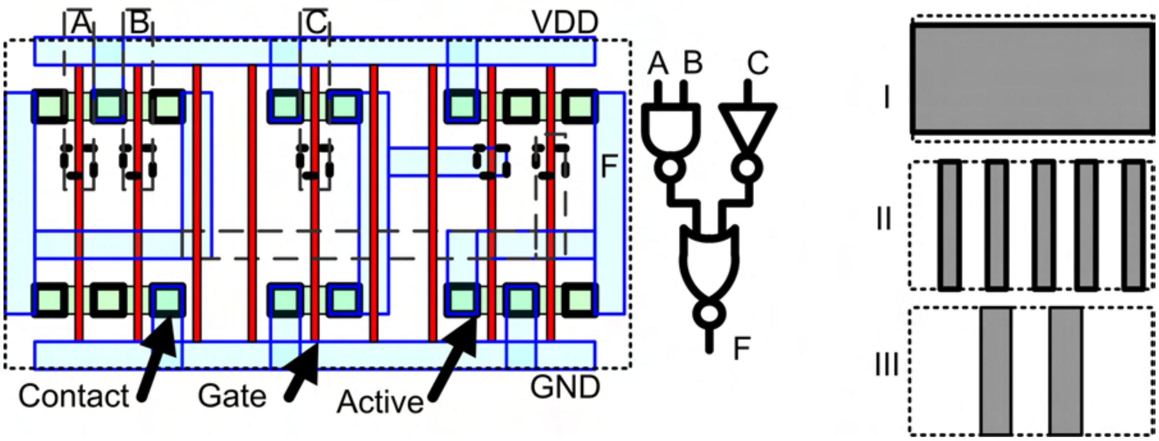

Figure 3.29. Negative spacer lithography is best suited for regular circuit layouts and can be adapted to arbitrary logic gates composed of standard cells. The active, gate, and contact layers of this circuit can be defined in seven spacer lithography steps using only three masks (right).

cut lines off the inverse of Mask III. The transistor gates are all defined with conventional spacer lithography (from Mask II) and separated with a trimming step (Mask I) at the edges of the circuit. If the pattern of Mask II is too small for photolithography, the pattern can be reproduced from larger features using the iterated process described in Section 3.3.3. Contact holes can be defined with a negative spacer lithography process using a combination of all three masks. In this manner, circuits can be integrated at sub-lithographic pitch. Since all of the critical device dimensions are spacer-defined, variability is expected to be minimized. The cost of an all-spacer integration is only in the extra processing required. Although four additional lithography steps are required relative to a photolithography process, no additional masks are needed.

3.5 Summary

The device techniques with the greatest potential to reduce variation in SRAM performance require the greatest departures from current CMOS technology. Multi-gate devices and spacer lithography processes address variations at the device parameter level. By reducing σ_{VT0} and σ_L, they reduce SRAM metric variations as well. The resulting yield improvements can significantly extend SRAM scaling, but at a high cost in processing and design.

The biggest yield improvement is expected from a transition to multi-gate devices. The use of undoped channels removes the leading cause of threshold voltage variation, random dopant fluctuation. The improved gate control of these devices also reduces DIBL, V_T rolloff, and off-state leakage, all of which improve SRAM performance or function. With triple-gate devices specifically, read and write yield improvements of over 2σ are expected, with no area or speed penalty (and in some cases, enhancement). Though good for SRAM, a transition to triple-gate devices will likely require the redesign of many logic circuits to satisfy device width constraints. Process improvements in lithography depth of focus, etching, and planarization may also be required.

Using spacer-defined devices will remove much of the variation associated with dimensions such as gate length and active width. V_T variations associated with CD problems such as LER are becoming worse with scaling. With multi-gate devices, tight CD control is expected to be even more important. Techniques such as negative and iterated spacer lithography allow for improved uniformity at sub-lithographic pitch, but extra processing is required. The greatest improvements in yield and area are expected from a combination of spacer processing and multi-gate devices, such as the $0.0512\mu\text{m}^2$ all-spacer FinFET SRAM described above.

The yield benefits of these device techniques are great enough that it is likely some form of them will eventually be adopted. However, the high costs and high risk will lead semiconductor companies to delay adoption of these techniques as long as possible. In the meantime, there is an opportunity for incremental solutions in the form of new peripheral circuit and cell designs. Circuit techniques can provide modest yield enhancements with moderate cost but low risk. In combination with process optimization, they may allow SRAM scaling to continue for the short term. More importantly though, they also complement the device techniques that have been presented in this chapter, addressing what variation will remain even if all of the device techniques here are implemented.

3.6 References

[1] T. Mizuno, J. Okamura, and A. Toriumi. Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's. *IEEE Trans. on Electron Devices*, pages 2216–2221, 1994.

[2] A. Asenov. Simulation of statistical variability in nanoscale MOSFETs. *VLSI Tech. Dig.*, pages 86–87, 2007.

[3] J.-P. Colinge. Reduction of floating substrate effect in thin-film SOI MOSFETs. *IEEE Electronics Letters*, pages 187–188, 1986.

[4] N. Lindert, Y.-K. Choi, L. Chang, E. Anderson, W. Lee, T.-J. King, J. Bokor, and C. Hu. Quasi-planar NMOS FinFETs with sub-100 nm gate lengths. *IEEE Device Research Conference*, pages 26–27, 2001.

[5] K. Hieda, F. Horiguchi, H. Watanabe, K. Sunouchi, I. Inoue, and T. Hamamoto. New effects of trench isolated transistor using side-wall gates. *IEEE Trans. on Elec. Dev.*, pages 736–739, 1987.

[6] J.P. Colinge, M.H. Gao, A. Romano-Rodriguez, H. Maes, and C. Claeys. Silicon-on-insulator 'gate-all-around device'. *IEEE International Electron Devices Meeting*, pages 595–598, 1990.

[7] K. Samsudin, B. Cheng, A.R. Brown, S. Roy, and A. Asenov. Integrating intrinsic parameter fluctuation description into BSIMSOI to forecast sub-15 nm UTB SOI based 6T SRAM operation. *Solid-State Electronics*, pages 86–93, 2006.

[8] V. P. Trivedi and J. G. Fossum. Scaling fully depleted SOI CMOS. *IEEE Trans. Electron Devices*, pages 2095–2103, 2003.

[9] S. Balasubramanian. Nanoscale thin-body MOSFET design and applications. *Doctoral Dissertation, University of California at Berkeley*, 2006.

[10] E. Nowak, T. Ludwig, I. Aller, J. Kedzierski, M. Leong, B. Rainey, M. Breitwisch, V. Gemhoefer, J. Keinert, and D. Fried. Scaling beyond the 65 nm node with FinFET-DGCMOS. *IEEE Custom Integrated Circuits Conference*, pages 339–342, 2003.

[11] T. Park, H. J. Cho, J. D. Choe, S. Y. Han, S.-M. Jug, J. H. Jeong, B. Y. Nam, O. I. Kwon, J. N. Han, H. S. Kang, M. C. Chae, G. S. Yeo, S. W. Lee, D. Y. Lee, D. Park, K. Kim, E. Yoon, and J. H. Lee. Static noise margin of the full DG-CMOS SRAM cell using bulk FinFETs (Omega MOSFETs). *IEEE International Electron Devices Meeting*, pages 27–30, 2003.

[12] R. V. Joshi, R. Q. Williams, E. Nowak, K. Kim, J. Beintner, T. Ludwig, I. Aller, and C. Chuang. FinFET SRAM for high-performance low-power applications. *European Solid-State Device Research Conference*, pages 211–214, 2004.

[13] Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. King, and B. Nikolic. FinFET-based SRAM design. *IEEE International Symposium on Low Power Electronics and Design*, pages 2–7, 2005.

[14] J. J. Kim, K. Kim, and C.-T. Chuang. Independent-gate controlled asymmetrical SRAM cells in double-gate MOSFET technology for improved read stability. *European Solid-State Circuits Research Conference*, pages 74–77, 2006.

[15] J.-T. Park and J.P. Colinge. Multiple-gate SOI MOSFETs: device design guidelines. *IEEE Trans. on Electron Devices*, pages 2222–2229, 2002.

[16] T. Park, S. Choi, D. H. Lee, J. R. Yoo, B. C. Lee, J. Y. Kim, C. G Lee, K. K. Chi, S. H. Hong, S. J. Hyun, Y. G. Shin, J. N. Han, I. S. Park, U. I. Chung, J. T. Moon, E. Yoon, and J. H. Lee. Fabrication of body-tied FinFETs (Omega MOSFETs) using bulk si wafers. *IEEE Symposium on VLSI Technology*, pages 135–136, 2003.

[17] X. Sun, Q. Lu, V. Moroz, H. Takeuchi, G. Gebara, J. Wetzel, S. Ikeda, C. Shin, and T.-J. King Liu. Tri-gate bulk MOSFET design for CMOS scaling to the end of the roadmap. *IEEE Electron Device Letters*, page in press, 2008.

[18] J. G. Fossum, L.-Q. Wang, J.-W. Yang, S.-H. Kim, and V. P. Trivedi. Pragmatic design of nanoscale multi-gate cmos. *IEEE International Electron Devices Meeting*, pages 613–616, 2004.

[19] L. Liu, K.L. Pey, and P. Foo. Hf wet etching of oxide after ion implantation. *IEEE Electron Devices Meeting*, pages 17–20, 1996.

[20] K. Takeuchi, R. Koh, and T. Mogami. A study of the threshold voltage variation for ultra-small bulk and SOI CMOS. *IEEE Transactions on Electron Devices*, pages 1995–2001, 2001.

[21] A. Carlson, X. Sun, C. Shin, and T.-J. King Liu. SRAM yield and performance enhancements with tri-gate bulk MOSFETs. To be published.

[22] Sentaurus Device Simulator, v.2006.06, Synopsys, Inc.

[23] D.A. Antoniadis and J.E.Chung. Physics and technology of ultra short channel MOSFET devices. *International Electron Devices Meeting*, pages 21–24, 1991.

[24] C. Shin, A. Carlson, X. Sun, K. Jeon, and T.-J. King Liu. Tri-gate bulk MOSFET design for improved robustness to random dopant fluctuations. To be published.

[25] A. Asenov. Random dopant induced threshold voltage lowering and fluctuations in sub-0.1μm MOSFETs: A 3-D atomistic simulation study. *IEEE Trans. Elec. Dev.*, pages 2505–2513, 1998.

[26] A. Balasinski, H. Gangala, V. Axelrad, and V. Boksha. A novel approach to simulate the effect of optical proximity on MOSFET parametric yield. *International Electron Devices Meeting*, pages 913–916, 1999.

[27] A. Bakri, M. Manaf, K. Wahab, and I. Ahmad. The characterization of KrF photoresists and the effect of different chromophore bulkiness on line edge roughness (LER) for submicron technology. *International Conference on Semiconductor Electronics*, pages 955–964, 2006.

[28] A. Dixit, K. G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, and K. De Meyer. Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: Role of line-edge-roughness. *IEEE International Electron Devices Meeting*, pages 709–712, 2006.

[29] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr. A 153MB-SRAM design with dynamic stability enhancement and leakage reduction in 45nm high-κ metal-gate CMOS technology. *International Solid-State Circuits Conference*, pages 376–377, 2008.

[30] S. Thompson, M. Alavi, R. Arghavani, A. Brand R. Bigwood, J. Brandenburg, B. Crew, V. Dubin, M. Hussein, P. Jacob, C. Kenyon, E. Lee, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, M. Prince, R. Schweinfurth, S. Sivakumar, P. Smith, M. Stettler, S. Tyagi, M. Wei, J. Xu, S. Yang, and M. Bohr. An enhanced 130nm generation logic technology featuring 60nm transistors optimized for high performance and low power at 0.7 - 1.4 V. *IEEE International Electron Devices Meeting*, pages 257–260, 2001.

[31] W.R. Hunter, T.C. Holloway, P.K. Chatterjee, and A.F. Tasch Jr. A new edge-defined approach for submicrometer MOSFET fabrication. *IEEE Elec. Dev. Let.*, pages 4–6, 1981.

[32] Y.-K. Choi, T.-J. King, and C. Hu. A spacer patterning technology for nanoscale cmos. *IEEE Trans. Elec. Dev.*, pages 436–441, 2002.

[33] A. Carlson and T.-J. King Liu. Negative and iterated spacer lithography processes for low variability and ultra-dense integration. *SPIE Advanced Lithography Conference*, page 6924.10, 2008.

[34] Y.-K. Choi, J. S. Lee, J. Zhu, G. A. Somorjai, L. P. Lee, and J. Bokor. Sublithographic nanofabrication technology for nanocatalysts and DNA chips. *J. Vac. Sci. Tech. B*, pages 2951–2955, 2003.

[35] R. Rooyackers, E. Augendre, B. Degroote, N. Collaert, A. Nackaerts, A. Dixit, T. Vandeweyer, B. Pawlak, M. Ercken, E. Kunnen, G. Dilliway, F. Leys, R. Loo, M. Jurczak, and S. Bisemans. Doubling or quadrupling MuGFET fin integration scheme with higher pattern fidelity, lower CD variation and higher layout efficiency. *IEEE International Electron Devices Meeting*, pages 993–996, 2006.

[36] M. LaPedus. Hurdles loom as foundry early adopters sprint toward 45 nm. *EETimes.com*, pages 1–4, Apr. 9, 2007.

Chapter 4

Designing for Variation in SRAM

4.1 Introduction

Although fundamentally its sources are at the device level, SRAM variability can also be addressed with circuit design. Circuit-based solutions have the advantages of low risk and low cost of development relative to changes in the device architectures or processes, which are often disruptive technologies. These advantages are compelling in a competitive industry in which product delays can result in a loss of market share and economic reasons may limit the future rate of scaling. The disadvantages of circuit-based solutions are that they do not address the $1/\sqrt{WL}$ increase in variability with scaling and their effectiveness does not scale. Instead they provide a one-shot boost to yield, with the possibility for additional robustness from supplementary circuit or device solutions. Nevertheless, their low cost and risk make circuit-based solutions attractive for short term yield enhancements, at least.

Many circuits for reducing SRAM variability have already been developed, with approaches ranging from introducing redundancy to modifying the biases, timings, or even the designs of the bitcell. Designs with redundancy use additional rows or columns of SRAM cells that can be programmed to replace a small number of failing cells. Error correction coding techniques, in which a small number of additional bits is used to verify and correct of a small section of memory, also fall in this category. SRAM arrays with redundancy are commonly used in commercial products, with an area overhead of up to 10% [1]. Other designs modify the biases on the bitlines, the wordline, V_{DD}, **GND**, or in the wells to facilitate read and/or write operations. For example, read stability can be enhanced by lowering wordline voltage or forward biasing the n-well around the **PU** devices. Arrays with dynamic n-well biasing have recently been incorporated into commercial products [2]. Still other designs modify the length of time for a read or write access to improve read stability or write-ability. A shorter read access time, for example, leverages the internal node capacitance of the bitcell to give a higher dynamic noise margin, even if SNM is small [3]. Finally, some memories use different bitcell designs to increase margins by adding extra transistors (e.g. 8-T SRAM [4]), introducing asymmetry [5], or making additional connections within the cell [6].

The approaches can be differentiated as *closed-loop*, which employs feedback to respond to measured variation, or *open-loop*, which improves robustness generally. Designs with redundancy are closed-loop insofar as they target cells with observed failure. On the other hand, designs which modify the bitcell are open-loop. Designs which modify biases or timing can be in either category. Among these, closed-loop designs can be expected to give better results if they are designed to optimize a tradeoff, such as that between read stability and write-ability. With this kind of feedback, it is necessary first to sense the device variations (e.g. in parameters such as W, L_G, or V_{T0}) or their net effect, and second to make the appropriate compensation via one of the methods described above. Often it is sufficient to sense variations only once, since the dominant sources like random dopant fluctuation and line edge roughness are invariant with time. These variations can thereafter be compensated with an open-loop implementation, for example by programming an optimal **WL** bias to balance read and write margins.

Closed-loop implementations require on-chip sensing of variations. In order to capture both systematic and random effects, circuits with SRAM-like layouts can be implemented in close proximity to the SRAM array. These circuits use wiring in the metal layers to focus on the pass-gate devices [7] or NMOS devices generally [8]. Takeda *et al.* use a "monitoring circuit" of one-and-a-half cells to sense variations that affect write-ability specifically. By providing on-chip measurements of variation, these designs can enable real time compensation. Off-chip sensing techniques can also be useful for characterizing the variability of a process. They have the added advantage of accuracy by measuring the cell or device currents from the actual SRAM cells in the array [9, 10, 11]; however, they are not practical for closed-loop compensation.

In addition to their benefits to yield, an important metric for these circuits is their layout area. Whether on the periphery of an array or in the bitcells themselves, every additional circuit expands the total area footprint of the memory. Circuits for reducing SRAM variability must therefore be compared in terms of area overhead. In addition, they must be compared against the simple alternative of slowing bitcell scaling.

This chapter presents several techniques to sense and compensate for SRAM variability. A variation sensor based on SRAM cells is introduced to sense how spatially-correlated variations affect the read / write balance of an array. A method for measuring statistical distributions of device parameters from within large SRAM arrays is presented. Finally, the advantages of feedback at the bitcell level are analyzed for the enabling case of independently-gated FinFETs, and designs to further enhance yield are proposed.

4.2 Systematic Variation Sensing

Process variations can be systematic or random in nature. Systematic variations, such as those caused by changes in critical-dimension bias or alignment, are spatially correlated, affecting multiple cells in close proximity. The spatial correlation can extend over several microns (within die) to several millimeters (die-to-die), or from wafer to wafer.

The variability of any particular SRAM cell is determined by a combination of many random and systematic variations at the device level. Both types of variation can affect device pairs in a common or differential (i.e., mismatch) mode. The relative importance

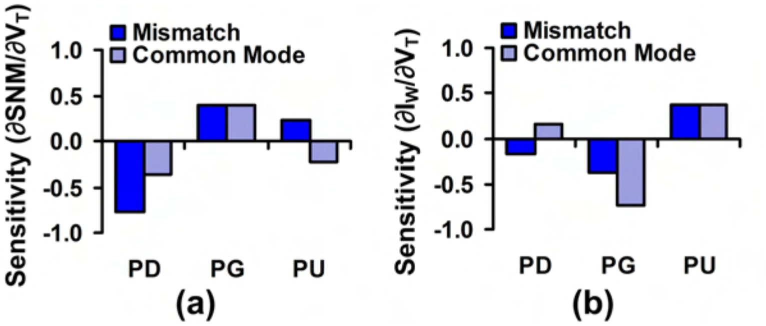

Figure 4.1. Normalized sensitivities to read SNM (a) and write-ability current (b), illustrating that both mismatch and common mode variations affect SRAM.

of common-mode and mismatch variations can be illustrated with simulated, normalized sensitivities of read stability and write-ability metrics to V_{T0} variation (Fig. 4.1). The sensitivities in Fig. 4.1 were simulated using measured I-V targets from padded-out SRAM cells in an early 45nm industrial process. Read SNM is much more sensitive to mismatch variations than common-mode variations in the **PD** devices, whereas it is equally sensitive to mismatch and common-mode variations in the **PG** devices, since each of these affects only one half of a cell (Fig. 4.1a). For write-ability, the sensitivities are more distributed, and are accentuated for common-mode variations in the **PG** devices (Fig. 4.1b) due to the complementary pull-down / pull-up behaviors of these devices on each side of the cell. Note that common-mode variations generally degrade either stability or write-ability via the read / write tradeoff, whereas mismatch variations always degrade one half of a cell. Control of both common-mode and mismatch variations is therefore important for SRAM robustness.

The analyses in chapters 2 and 3 focused on random variations for several reasons. The magnitude of random variations has been increasing with continued technology scaling and is beginning to reach prohibitive levels. Unlike systematic variations, random variations cannot be compensated by a process adjustment. Systematic variations are harder to model, since they are not independent and can arise from complex sources.

Systematic variations can cause both common-mode and differential variations in the cell performance because they can arise from multiple, opposing sources. For example, a systematic bias in transistor gate lengths would be a source of common-mode variation,

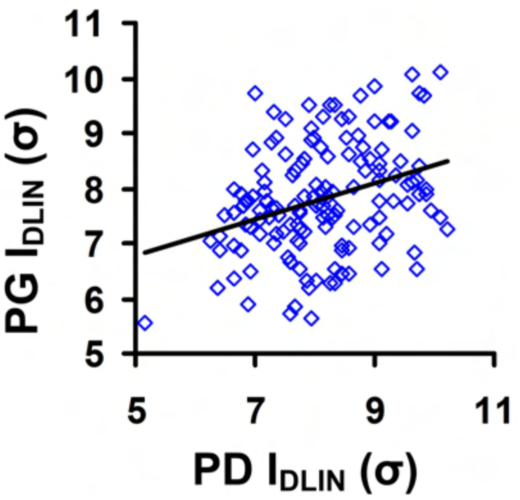

Figure 4.2. I_{DLIN} measurements of adjacent **PG** and **PD** devices are weakly correlated ($R^2 = 0.11$). The standard deviation of currents can be decomposed into systematic (σ_s) and random (σ_r) components, with systematic variations accounting for approximately 10% of total **PG** I_{DLIN} variability.

whereas a gate misalignment would cause a mismatch. The total variation can be decomposed into a net systematic component (σ_s) and a random component (σ_r). Fig. 4.2 illustrates the correlation between adjacent **PD** and **PG** I_{DLIN} ($V_{DS} = 0.1$V, $V_{GS} = 1.0$V) measured from 144 cells. The correlation of the currents is $R^2 = 0.11$, suggesting that approximately 11% of the total variation is due to spatially correlated sources. Variations from a perfectly random source alone would be expected to give a correlation less than this with 91% confidence. Among all device pairings, the correlation between **PG** and **PD** I_{DLIN} was the largest, indicating that the process has a very low within-die systematic variability. Greater systematic variations are expected at the die-to-die and wafer-to-wafer levels, however.

In order to detect spatially correlated variations, small sensor circuits can be placed on the periphery of an SRAM array [8, 7, 12], using partial cell layouts to replicate the environment of the actual SRAM cells. For example, Agarwal *et al.* use a mini-array of SRAM-sized devices to enable I-V data collection [8]. Off-chip sweep and measurement equipment can be used to quantify both random and systematic variations, which can then be used to estimate yield. Other designs combine the sensing of systematic variations with a circuit to correct them using cell bias. The information from the sensors can be used to compensate average read or write margins of the arrays through V_{DD} adjustments, well-, bitline- or WL-biasing. Ohbayashi et al. use several "replica access transistors" with an identical layout to PG devices in the cell [7]. These transistors oppose a wide PMOS driver to set the WL bias voltage (V_{WL}) for cell access. The circuit thereby compensates for systematic NMOS variations by adjusting V_{GS} on the **PG** transistors, which increases robustness to read upset but degrades write-ability. Additional circuitry is used to recover the write-ability to a desired minimum threshold. Takeda et al. use a "monitoring circuit" of one-and-a-half cells to determine the minimum V_{WL} or maximum V_{DD} to ensure write-ability [12]: the internal nodes of a monitor cell are connected to an op-amp, which sets $V_{DD} - V_{WL}$ to force the cell into a meta-stable state (with equal internal node voltages), which is the minimum DC condition for a successful write. This approach can increase robustness during read well beyond what is practically required, with an excessive penalty to write access time.

Cells in the same row must simultaneously be both writeable and robust to half-select upset. There is a well-understood tradeoff between the write access time and stability during a half-select condition, which at DC can be measured by the read static noise margin. Increasing the strength of the **PG** devices improves write-ability but degrades read SNM. Increasing the strength of the **PU** devices tends to produce the opposite effect. Due to this read / write tradeoff, optimizing for write-ability impairs the half-select stability, and vice-versa. An improved approach to correction for systematic variations that optimizes this tradeoff can maximize overall SRAM yield.

4.2.1 Sensor Design

The systematic process variation can be expected to degrade either write-ability or cell stability during a half-select. Fig. 4.3 illustrates a variation sensor circuit that restores the balance using the read / write tradeoff. The sensor comprises half-cells configured for a

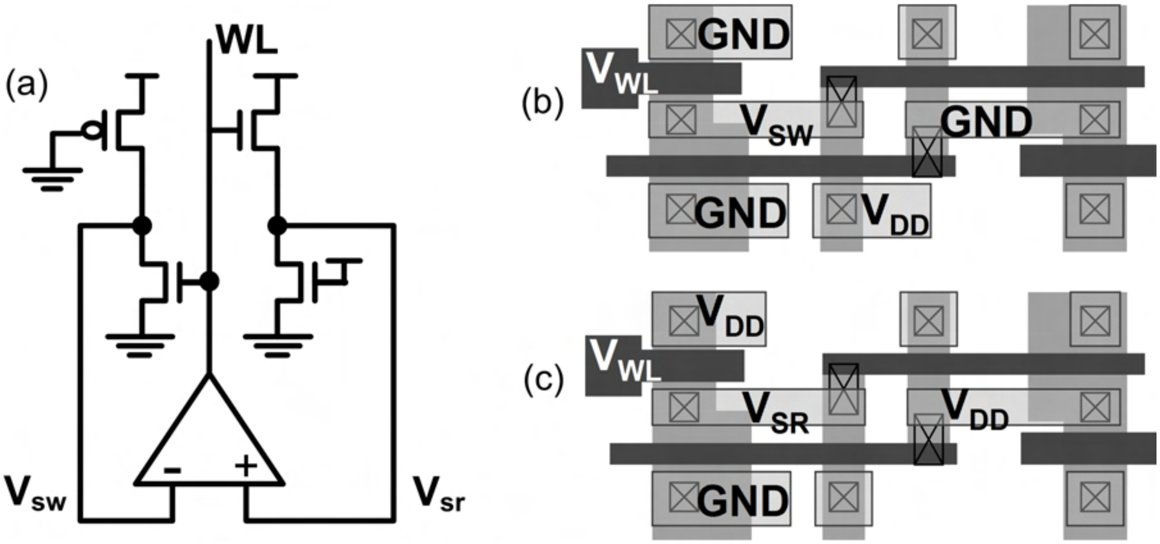

Figure 4.3. Multiple copies of cells with identical SRAM layouts through the first metal layer are configured for worst case write (a, left side; layout: b) or worst case read (a, right side; layout: c). The voltages at the internal nodes V_{sw} for write and V_{sr} for read are negatively correlated with cell read stability and write-ability. The op-amp changes V_{WL} to optimize the read / write tradeoff for systematic variations.

worst-case read or a worst-case write. Actual SRAM bitcell layouts are used up through the first metal layer to ensure maximum sensitivity to layout-sensitive variations. The worst-case condition for writing a cell consists of a voltage divider with the **PG** device contesting a fully-on **PU** device to bring the internal node voltage, V_{sw}, low. In the layout, this can be achieved (for example, on the **CH** side) by connecting **BL** and **CL** to **GND**, with V_{DD} at the source of the appropriate **PU** device, and the other nodes floating (Fig. 4.3b). Similar connections can be made to configure the cell for a worst-case write on the **CL** side. The worst-case read condition for a cell consists of a resistive voltage divider with a fully-on **PD** contesting the adjacent **PG** device to bring the internal node voltage, V_{sr}, to a low value. In the layout, this can be achieved by connecting **BL** and **CL** to V_{DD} and by connecting V_{SS} to **GND** on the **CH** side of the cell (Fig. 4.3c). The rest of the nodes are left floating. In all cases, the gate of the **PG** is connected to **WL**, and each sensor half-cell is the size of one SRAM bitcell.

The internal node voltages, V_{sr} and V_{sw}, can be used to estimate the read stability and write-ability of a set of cells. Lower V_{sr} corresponds to less read disturb and therefore higher

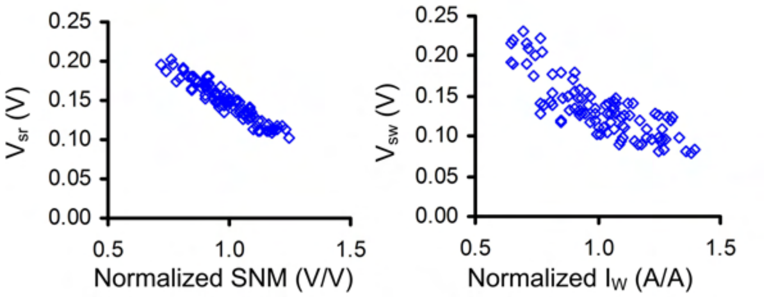

Figure 4.4. The outputs of the worst-case read and write sensor cells, V_{sr} and V_{sw}, are negatively correlated with SNM and I_w, respectively. The read / write tradeoff for a set of cells can be estimated with minimal area overhead using these metrics.

read stability. Lower V_{sw} corresponds to an easier and faster write. Fig. 4.4 illustrates good correlation between these voltages and read SNM or I_W, as determined by Monte Carlo simulation. The correlations in Fig. 4.4 can be improved by including secondary contributions from the other four transistors in the SRAM cell, but this requires additional connections that increase overall sensor area.

In both read and write cases, the strength of the **PG** transistor is determined by V_{WL}. Wordline biasing was chosen due to its relatively large impact on the write / read (half-select) tradeoff; however, other biasing or timing approaches could also be used. Raising V_{WL} increases the **PG** strength, lowering V_{sw} but raising V_{sr}. An op-amp is used to find the optimal wordline bias ${V_{WL}}^*$ for which the maximum of V_{sr} and V_{sw} is minimized (Fig. 4.5). Because of the read / write tradeoff, the minimum is achieved when $V_{sr} = V_{sw}$. Depending on the requirements of the application, ${V_{WL}}^*$ may be chosen to be a fixed offset away from this point without affecting the sensor's response to systematic variations.

In the presence of variations, the sensor modifies V_{WL} to maintain $V_{sr} = V_{sw}$. Fig. 4.6 illustrates the simulated response of a sensor to variations in the **PG** gate length. Increasing gate length weakens the **PG** transistor, improving read stability, and raising V_{sw}. V_{sr} and V_{sw} are plotted with the op-amp output connected to V_{WL} in feedback vs. with V_{DD} connected to **WL**. The sensor correctly lowers the maximum of V_{sr} and V_{sw} at all points but one, where the optimal ${V_{WL}}^* = V_{DD}$.

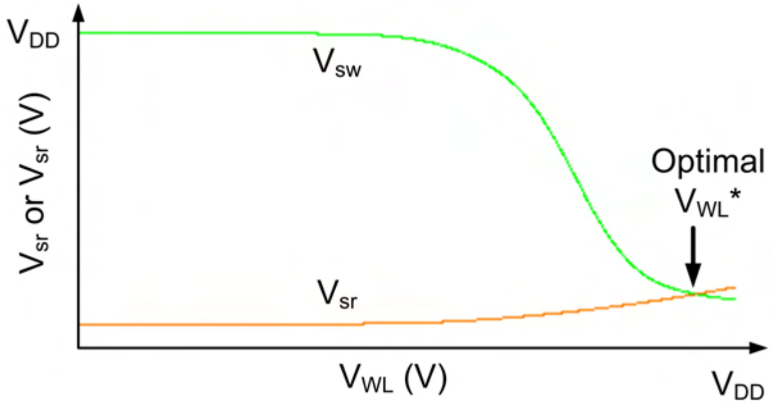

Figure 4.5. The optimal V_{WL}^* is that which minimizes the maximum of V_{sr} and V_{sw}. Because of the read / write tradeoff, this point is where the two voltages are equal.

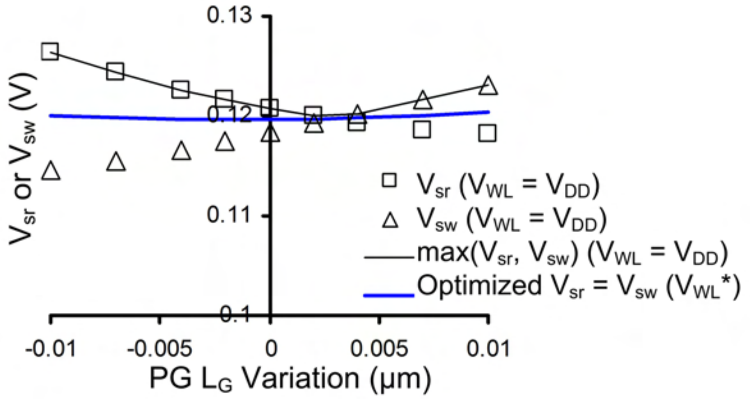

Figure 4.6. The optimal V_{WL}^* is adjusted to minimize the maximum of V_{sr} and V_{sw} in the presence of systematic variations, such as on **PG** L_G.

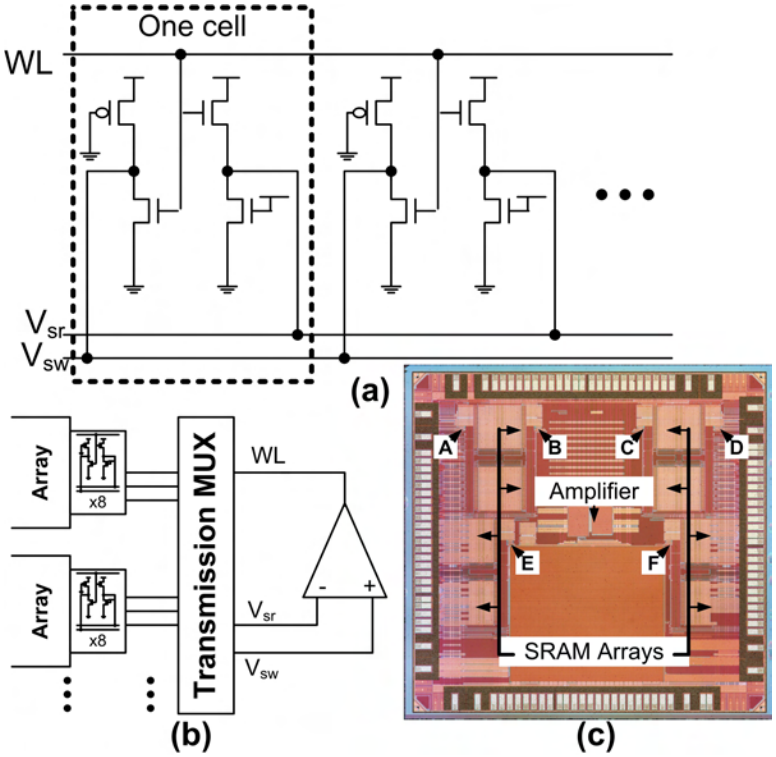

Figure 4.7. Variation sensors consisting of 16 half-cells are implemented on an SRAM testchip in an industrial 45nm process (a). The sensors are situated in close proximity to SRAM arrays and share a single amplifier, connected through a transmission multiplexer (b). Die photo of the testchip (c).

Each sensor can comprise multiple, parallel copies of half-cells. The cells are connected in parallel with the outputs V_{sr} tied together for the read half-cells, and V_{sw} tied together for the write half-cells. Increasing the number of parallel half-cells reduces the variability due to random sources by a factor of $1/\sqrt{N}$. The half-cells can optionally include different orientations to control for systematic variations associated therewith.

4.2.2 Experimental Results

The variation sensor is implemented in silicon in an early industrial 45nm process, as illustrated in Fig. 4.7. Six sensors are placed near SRAM arrays across a 2mm × 2mm die. Each sensor consists of sixteen half-cells: eight for the worst-case read and eight for the worst-case write. All cell orientations are included. The input and outputs of each sensor (V_{WL}, V_{sr}, and V_{sw}) are connected via a transmission multiplexer to one shared PMOS-

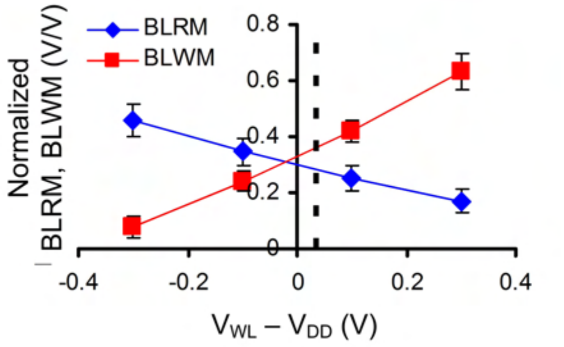

Figure 4.8. BLRM, a measure of read stability, decreases with increasing WL voltage, while BLWM, a measure of write-ability, increases for several SRAM cells in a dense array. Error bars indicate +/- one standard deviation. The optimal **WL** voltage of the nearest variation sensor is indicated by the dotted line.

input amplifier. In this proof-of-concept design, a simple 1-stage folded-cascode amplifier with a gain of more than 60dB is implemented in a 9000 μm^2 area. Large 1.8V, thick-oxide transistors were used to ensure a high gain and small 3σ offset of 2.4mV. In a mature process, when the matching data is known, a much smaller, thin-oxide amplifier could be used. V_{WL} is also connected to an external pin for data collection.

Fig. 4.8 illustrates the average response of several SRAM cells to different V_{WL}. Because only the bitlines and wordlines are externally accessible in these cells, the bitline read margin (BLRM) and bitline write margin (BLWM) are used to measure read stability and write-ability. BLRM is defined as V_{DD} minus the smallest cell supply voltage that retains the cell state with V_{BL}, $V_{\overline{BL}}$, and $V_{WL} = V_{DD}$. BLWM is defined as the highest V_{BL} that flips the cell state with $V_{WL} = V_{\overline{BL}} = V_{DD}$. These metrics have been shown to correlate well with read SNM and I_W, respectively [10]. With increasing V_{WL}, BLRM decreases and BLWM increases, similar to the expected trend from Fig. 4.5. The $V_{WL}{}^*$ indicated by the variation sensor is near the crossover point where BLRM = BLWM. A small offset can be expected since the metrics are different from the V_{sr} and V_{sw} equalized by the op-amp.

V_{DD} and substrate biases can also be used to skew the read / write tradeoff. With $V_{WL} = V_{DD}$, increasing V_{DD} improves write-ability faster than read stability (Fig. 4.9a). A lower V_{WL} can restore the balance, so the variation sensor output generates a decreasing

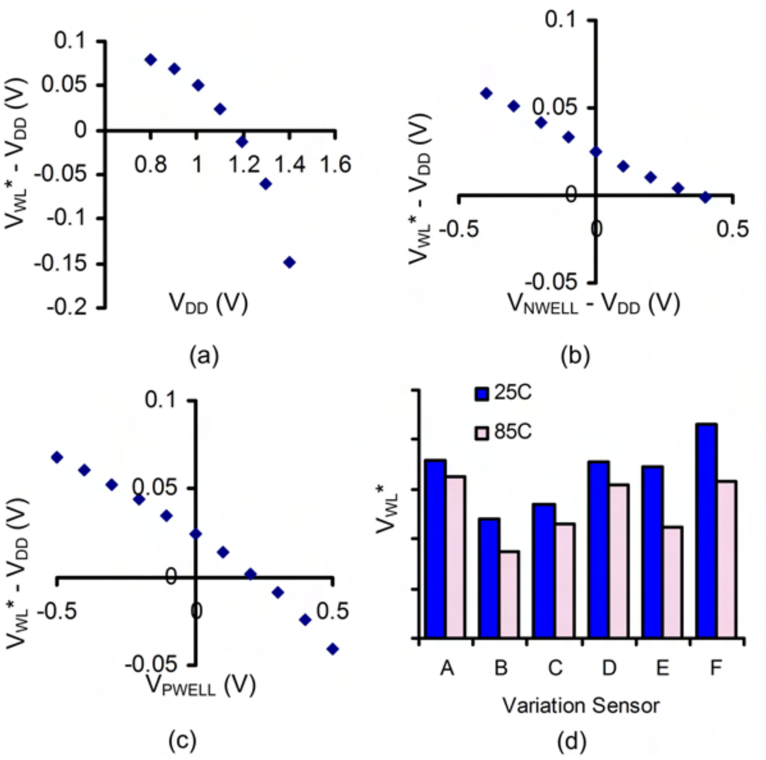

Figure 4.9. Increasing V_{DD} (a), n-well bias (b), p-well bias (c), and temperature (d) all have the effect of favoring write-ability. A lower ${V_{WL}}^*$ can be used to restore the balance.

${V_{WL}}^* - V_{DD}$. The read / write tradeoff is highly sensitive to V_{DD} variations, with an average of 40 mV of ${V_{WL}}^*$ required to compensate for a 100 mV change in V_{DD}. With increasing n-well bias, the PMOS devices become weaker, improving the write-ability. This effect is similar to that produced by burn-in phenomena such as negative bias temperature instability (NBTI), which increases PMOS V_{T0} over time and was shown in Section 2.5 to lower SRAM yield. The variation sensor appropriately decreases ${V_{WL}}^*$ in response to the rise in PMOS VT (Fig. 4.9b). With increasing p-well bias, the NMOS devices become forward-biased, increasing their drive currents, and improving the write-ability. ${V_{WL}}^*$ is observed to decrease accordingly (Fig. 4.9c). An average decrease in ${V_{WL}}^*$ with temperature of $\partial {V_{WL}}^*/\partial T = -0.2$ mV/K was also observed (Fig. 4.9d).

In order to evaluate across-chip variations, 512 cells from each of six arrays at different locations on the die are measured. BLRM and BLWM are measured for each cell at $V_{WL} = V_{DD} - 0.1$V and $V_{DD} + 0.2$V. A linear relationship between BLRM, BLWM, and V_{WL} is

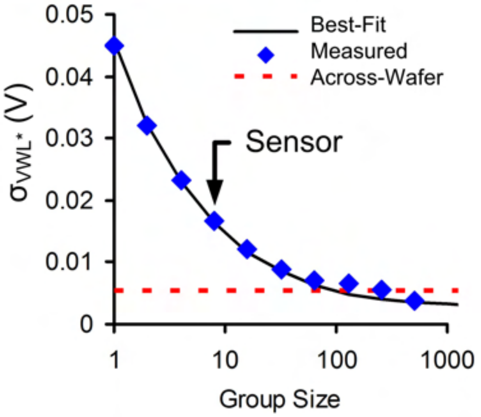

Figure 4.10. Measurements of groups of N SRAM cells show a decreasing random variation component as N increases. The least squares fit from theory (eqn. 4.1) indicates σ_s = 3mV, far below the detection limit of the 16 half-cell sensors. The across-wafer signal is also low (4mV).

assumed based on the measurements illustrated in Fig. 4.8. The V_{WL} where BLRM = BLWM is then calculated for each cell. Because of variability, there is a distribution of V_{WL} from each array. The V_{WL} are normally distributed with a standard deviation of 47 mV that comprises a random (σ_r) and systematic (σ_s) component. To isolate the systematic variation, groups of N cells can be averaged. The calculated V_{WL} are then expected to be normally distributed with a standard deviation of:

$$\sigma_{VWL} = \sqrt{\frac{\sigma_r^2}{N} + \sigma_s^2} \tag{4.1}$$

Fig. 4.10 illustrates the measured σ_{VWL} for different N among all six arrays. The least squares fit curve of Eqn. 4.1 is also shown, with σ_r = 46mV and σ_s = 3mV. This corresponds to a very low level of across-chip variation, far below the 17 mV that the implemented sensors are able to measure with only N = 8 cells. The measured value of σ_{VWL*} from 42 sensors on seven chips (with die-to-die and wafer-to-wafer systematic variations subtracted out) is 14 mV, close to the expected value. A large sensor of $N > 512$ cells would be required to detect systematic variations at σ_s = 3mV; however, such a sensor would be impractical since this level is too small to affect yield.

With simultaneous process, voltage, and temperature variations, however, a larger spread in the data is measured. At each point in Fig. 4.11, the average of all six variation

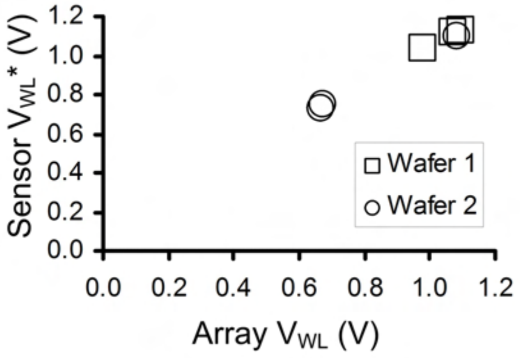

Figure 4.11. Chip-to-chip correlation of sensor ${V_{WL}}^*$ and optimal V_{WL} as measured from arrays. Each data point represents an average of six variation sensors and over 256 array cells. A range of chips, V_{DD}, substrate biases, and temperatures were used to verify operation under various operating conditions.

sensors on a chip is plotted against the average optimal V_{WL} (at BLRM = BLWM) from the arrays. To reduce noise, only chips with over 256 measured cells are plotted. The sensor ${V_{WL}}^*$ has a 50mV offset, but otherwise tracks the optimal value within 30mV over a 500mV range. Since each chip has its own op-amp, an additional source of variability in the sensor output is expected.

The variation sensor can compensate for high levels of systematic variation, where a reduction in σ_s is expected to improve yield. The sensors should contain at least $N > \sigma_r^2/\sigma_s^2$ cells to ensure a strong signal. The separation between multiple sensors on a chip determines the minimum distance over which spatially correlated variations can be detected. Sensors placed 400μm apart, for example, will be unable to detect variations that are correlated within 40μm, even if they are large in magnitude. In the test arrays, measurements for cells within ranges of 10 - 200μm exhibited a very low σ_s in this process, however, so only one sensor is required for these ranges.

4.3 Device Characterization from SRAM Measurements

One disadvantage of relying on peripheral sensors to characterize variations is that a significant layout area can be required to minimize their sensitivity to random variation or to provide statistics with high confidence. With SRAM yield requirements exceeding five sigma, statistical data is necessary to make accurate yield projections for new designs or processes. Throughout this work, the yield projections rely on an assumption of normality in the distributions of device parameters such as V_{T0} or L_G. For small amounts of variation, this assumption can be experimentally confirmed; however, some reported SRAM measurements show non-Gaussian behavior in the tails of the distribution, beyond the three sigma level [13]. For robust SRAM design, it is desirable to characterize the statistics of device parameters in SRAM layouts. These statistics can then inform process modifications or the next stage of design.

There are several advantages to gathering this data directly from the SRAM arrays themselves. The area requirements of sensor cells or separate device arrays [8] are greatly reduced by using the SRAM bitcells themselves. There are no effects from using different layout environments or positions in the die. Finally, the SRAM cells have a demonstrably high sensitivity to variation. The challenge is that there are many variable device parameters within a bitcell and relatively few external connections through which to characterize them.

SRAM arrays are already measured for statistics on performance and yield. So-called “schmoo” plots are common for graphing yield as a function of voltages and/or frequency. Measurements on the array can range from simple logical tests, such as those commonly used by built-in self-test (BIST) circuits, to current measurements through the bitlines. For example, Yu *et al.* use read current measurements to determine cell biases, thereby reducing die-to-die variations [9]. Current and voltage measurements have also been used to measure read stability and write-ability from large arrays [10]. With significant **PG** overdrive, the I-V characteristics of cell transistors can be measured directly [11].

I-V measurement of the cell transistors allows for direct electrical measurement of some device parameters, such as V_{T0}, and extraction of others, such as L_G (approximated from DIBL) and W (combined with variations in gate oxide thickness and mobility). There are some drawbacks, however, to measuring the I-V characteristics directly. First, leakage paths in the array make it difficult to measure currents below a few μA/μm. Threshold voltage must therefore be measured at a current level already in strong inversion or be extracted from high current measurements using I_D-V_G equations. Second, very high V_{WL} is required to minimize uncertainty in the internal node voltages for **PD** and **PU** measurements. This causes significant stress on the **PG** gate oxide, lowering the lifetime of the transistor. Third, there is a significant area cost associated with circuits that measure current. Finally, the device parameters must still be extracted from the I-V data. In spite of the significant disadvantages, this approach nevertheless enables high counts of data collection for determining the statistics of SRAM device variation [11].

These disadvantages can be eliminated with an approach using only measurements of SRAM metrics and knowledge of their sensitivities to device parameter variations. The approach is based on the pseudoinverse technique for solving systems of linear equations. Given a set of n device parameters $\vec{x}$, represented in vector form, and a set of m SRAM metrics $\vec{b}$, there exists an $m \times n$ matrix A such that the entries of A contain the sensitivities of the different metrics to the different parameters:

$$A_{ij} = \frac{\partial b_i}{\partial x_j} \tag{4.2}$$

In general, if $m \neq n$ or the parameters or metrics of $\vec{x}$ or $\vec{b}$ are not linearly independent, A is not invertible. In these cases, the device parameters $\vec{x}$ can be estimated with the pseudoinverse of A:

$$\vec{x} = \left(A^T A\right)^{-1} A^T \vec{b} \tag{4.3}$$

Eqn. 4.3 gives the $\vec{x}$ that minimizes $||A\vec{x} - \vec{b}||$ or, in the case of degenerate solutions, the one that minimizes $||\vec{x}||^2$. In practice, there is always error in the sensitivities and metric measurements, which results in a subspace of possible $\vec{x}$ meeting a minimum threshold of fit. For most cases though, the subspace can be made small and a good approximation of the actual $\vec{x}$ is obtained. This approach relies on two key assumptions: first, that the measured SRAM metrics can be represented as a linear combination of device parameters, and second, that the device parameters are linearly independent.

The first assumption is verified on 144 padded-out SRAM cells, in which the internal nodes can be directly accessed via large transmission gate multiplexers. I-V characteristics

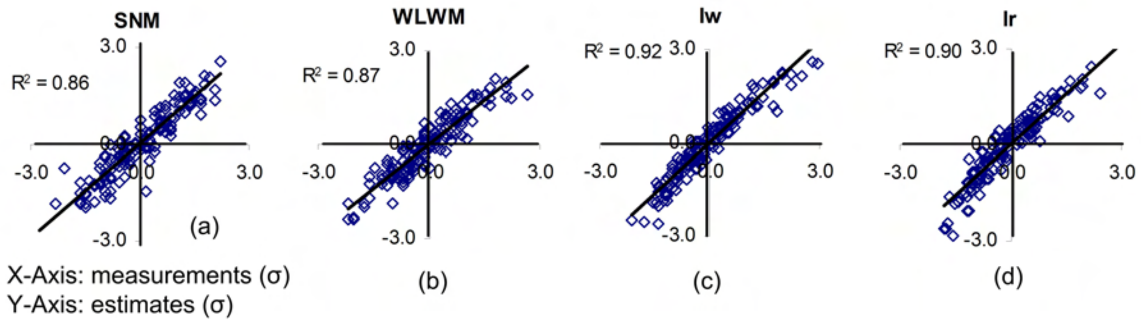

Figure 4.12. Linear combinations of device parameters can be used to estimate SRAM metrics such as SNM (a), WLWM (b), I_W (c), and I_r (d). Good correlations to the actual measured metrics are seen in all cases, using only two I-V targets per device.

for every transistor in every cell are measured, and the I-V targets V_{TLIN} and I_{DSAT} are used as device parameters. I_{DSAT} represents a combination of W, L_G, and V_{T0} variations–in addition to other possible sources of variation (such as in mobility or gate capacitance)–whereas V_{TLIN} is determined primarily by V_{T0}. SRAM metrics such as SNM, WLWM, I_w, and the DC read current I_r are measured from the same cells. The metrics SNM and I_w cannot be measured from the wordlines and bitlines alone, but they are used as proxies in this analysis for the bitline metrics BLRM and BLWM, which are not measurable in this padded-out implementation. Using the sensitivities as coefficients, these metrics can be estimated with a linear combination of the device parameters. Fig. 4.12 illustrates the correlation of these estimates with the actual measurements from the cell. Good correlations are seen to all of the metrics, verifying that they can be represented with this approach. The error can be attributed to device variations that are not captured by V_{TLIN} and I_{DSAT} alone.

The second assumption of linear independence in the device parameters can also be verified from measurements. Within a device, I_{DSAT} is strongly correlated with V_{TLIN} ($R^2 = 0.7$), but also includes a significant and detectable independent component. Among different devices, the parameters are also expected to be independent, in part due to the high level of random variation observed in section 4.2. A low measured correlation between parameters supports this assumption (Table 4.1). The highest correlations of 0.10 are

Table 4.1. Correlations of Device V_{TLIN} (units of σ_{VT})

	V_{T1}	V_{T2}	V_{T3}	V_{T4}	V_{T5}	V_{T6}
V_{T1}	1.00	0.10	0.10	0.06	0.06	-0.02
V_{T2}	0.10	1.00	0.08	0.07	0.07	0.04
V_{T3}	0.10	0.08	1.00	0.00	-0.05	-0.01
V_{T4}	0.06	0.04	0.00	1.00	-0.02	0.07
V_{T5}	0.06	0.07	-0.05	-0.02	1.00	0.06
V_{T6}	-0.02	0.07	-0.01	0.07	0.06	1.00

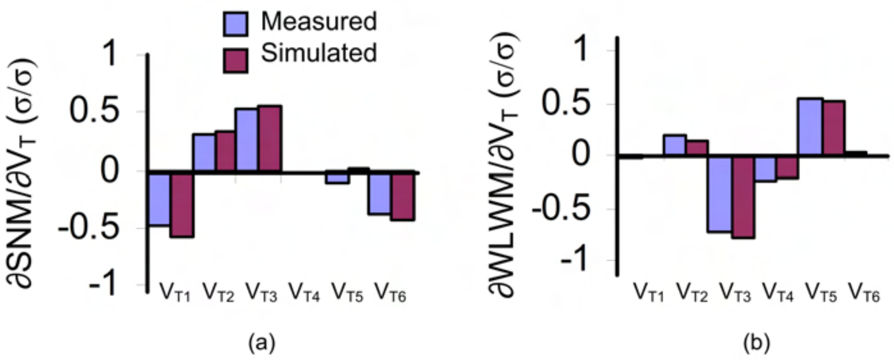

Figure 4.13. Sensitivities (in units of σ_{SNM} or σ_{WLWM} per σ_{VT}) extracted from simulation and a small number of padded-out SRAM cells show good agreement.

observed among NMOS V_{TLIN}.

With linearly independent and uncorrelated device parameters, it is easy to determine the sensitivity matrix A. In this case the sensitivities can be extracted from simulation or measurements of a small sample of padded-out cells, such as those in a scribe-line macro or a mini-array. From simulation, the sensitivities can be determined by measuring the change in an SRAM metric for a positive and negative perturbation in a single parameter. From measurements, the sensitivities can be extracted from the slope of a least-squares fit to the data. Fig. 4.13 illustrates the sensitivities to SNM and WLWM with good agreement between these methods. The presence of measurable SRAM cells on the wafer are expected to provide the more accurate sensitivities for characterizing device parameters. In the case where the device parameters are chosen to be linearly independent yet partly correlated, a set of uncorrelated device parameters can be chosen by subtracting out the projection onto

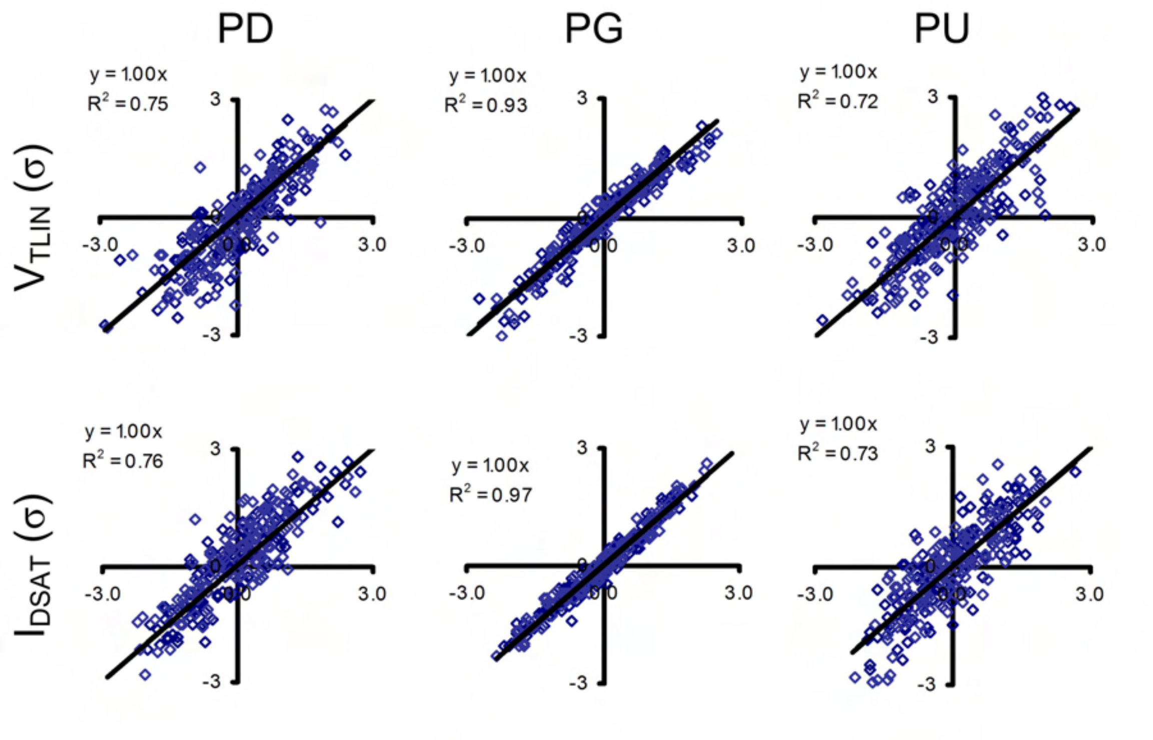

Figure 4.14. Device parameters can be extracted from a combination of SRAM current and voltage metrics at different V_{DD}. X-Axis: measured parameters, Y-Axis: extracted from SRAM model

the independent parameters. Alternatively, partly correlated parameters can be used for the extraction if the sensitivities in A are similarly adjusted.

Extraction of device parameters is first demonstrated using a combination of current and voltage metrics (SNM, WLWM, I_W, and I_r) at $V_{DD} = 0.6, 1.0,$ and 1.3V. Fig. 4.14 illustrates a good correlation between the extracted V_{TLIN} and I_{DSAT} from the SRAM metrics and the actual I-V targets measured from the devices. A total of 144 cells are plotted. The strongest correlations are to the parameters of the **PG** device, since it is directly connected to the bitlines and wordline. There is more error in the **PD** and **PU** extractions, but overall the average parameter error is low. Fig. 4.15 presents a histogram of the average parameter error, as determined by first calculating the absolute value of the error between extraction and measurement for each parameter in each device and, secondly, averaging among all the devices in the cell. The average is approximately 0.33σ, and three quarters of the cells have less than 0.4σ. Fig. 4.16a illustrates the I_{DSAT} and V_{TLIN} from

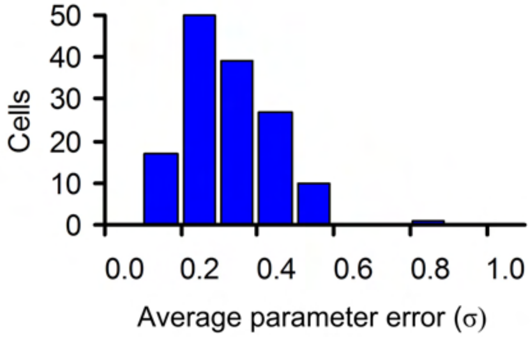

Figure 4.15. The average parameter error is approximately 0.33σ and less than 0.4σ for 75% of the cells .

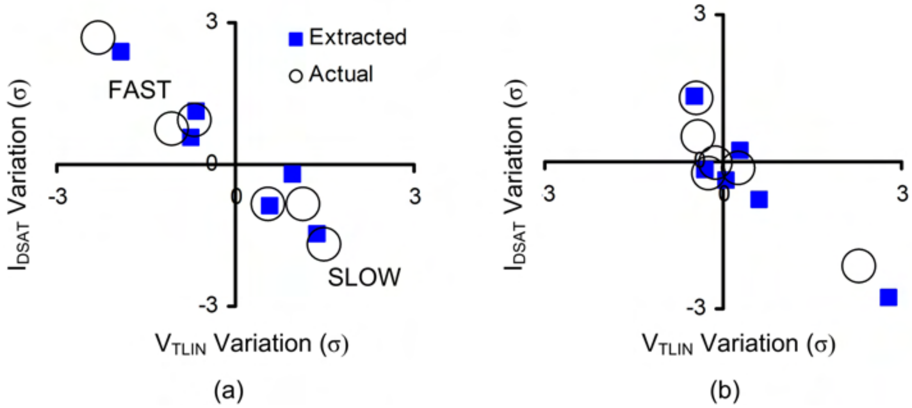

Figure 4.16. Device parameter extraction can indicate which devices are fast or slow, even if multiple devices exhibit variation in the cell (a). The approach also works for cells with a single outlier device (b).

extraction and measurement for a single cell. Devices in the top-left quadrant are relatively fast, with low V_{TLIN} and high I_{DSAT}. Devices in the bottom-right quadrant are relatively slow. The plot shows a good fit between the extraction and the actual measured parameters, even though the cell has a moderately high level of variation. This approach can also be used to extract parameter variation from cells with only one outlier device (Fig. 4.16b).

Even if the area and time costs of current measurements are too high, significant information can be extracted using voltage metrics alone. Fig. 4.17 illustrates the correlations between extracted and measured device parameters for the same 144 cells when only the voltage metrics SNM and WLWM are used. Measurements are taken at

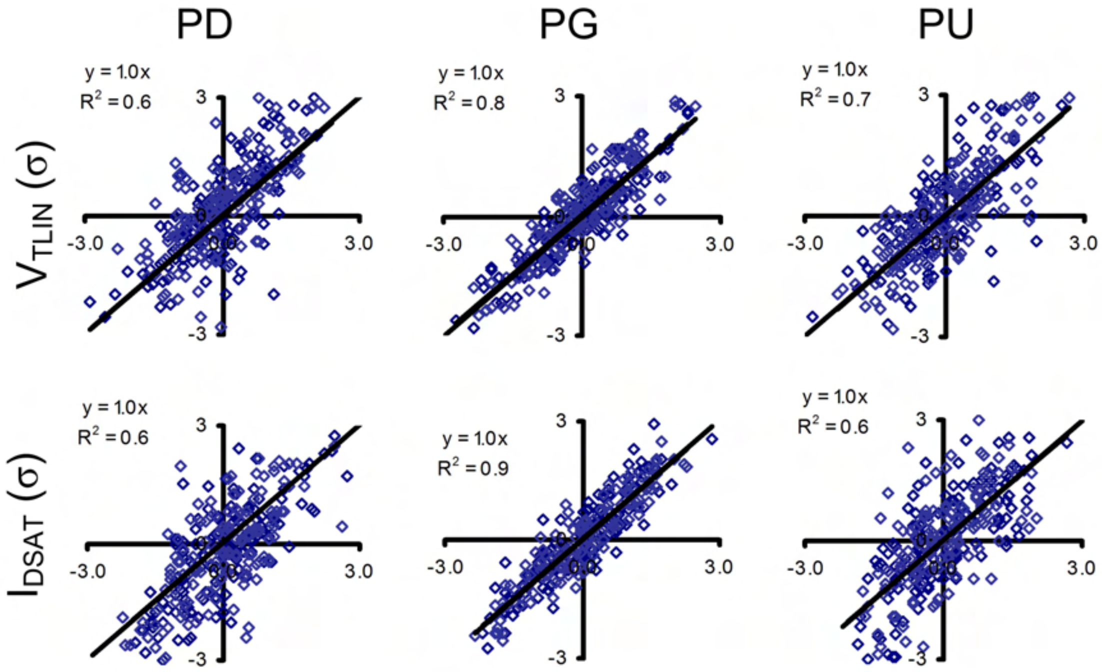

Figure 4.17. Even if only SRAM voltage metrics are used, device parameters can still be extracted with reasonably good accuracy. X-Axis: measured parameters, Y-Axis: extracted from SRAM model

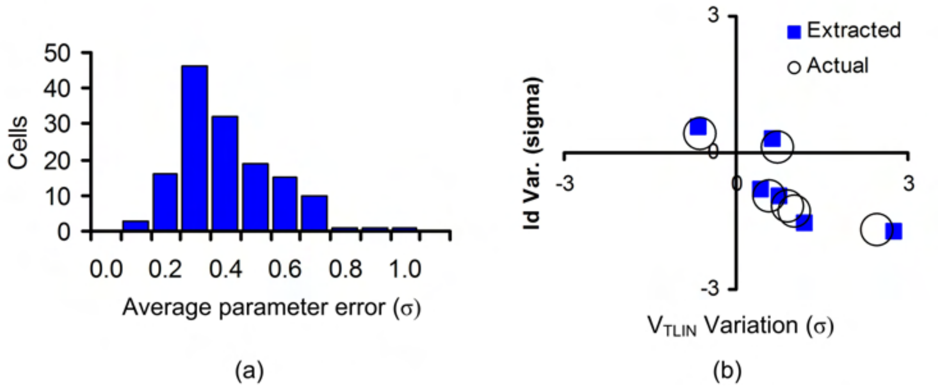

Figure 4.18. The average parameter error is 0.1σ larger when the device parameters are extracted using only voltage metrics (a); however, the relative variability for all six devices in a cell can still be obtained accurately (b).

$V_{DD} = 0.6, 1.0$, and 1.3V and also with forward body biasing of 0.4V on the n-well or the p-well at $V_{DD} = 0.6$V. The R^2 of the correlations is about 0.1 lower without the current measurements, but the average parameter error is still relatively low. Fig. 4.18a presents a histogram for the average parameter error in this case. The average is 0.5σ, but the median is 0.4σ. A third example cell is also shown with reasonably good fit (Fig. 4.18b). It is therefore feasible to extract device parameters from voltage metrics alone, eliminating the large area requirements of current measurement circuits.

The approach is demonstrated on 512 cells in a densely packed array, with the internal nodes inaccessible. The SRAM metrics BLRM and BLWM were measured at $V_{DD} = 1.0$V and 1.3V. Fig. 4.19 illustrates a spatial variation plot and the measured distribution for **PG** V_{TLIN}. Each square in the spatial variation plot represents the average of eight cells. The amount of spatially-correlated variation is very low and is consistent with the findings of section 4.2. The distribution is Gaussian shaped and similar to that from measurements of 512 **PG** devices from padded-out cells with the same cell layout. The standard deviations are comparable; however, the expectation was for a slightly larger σ in the extracted distribution due to the parameter error. The comparable σ between the distributions in this case is most likely due to an underestimation of the sensitivities, which happens to cancel the additional parameter error. The sensitivities to BLRM and BLWM were estimated only

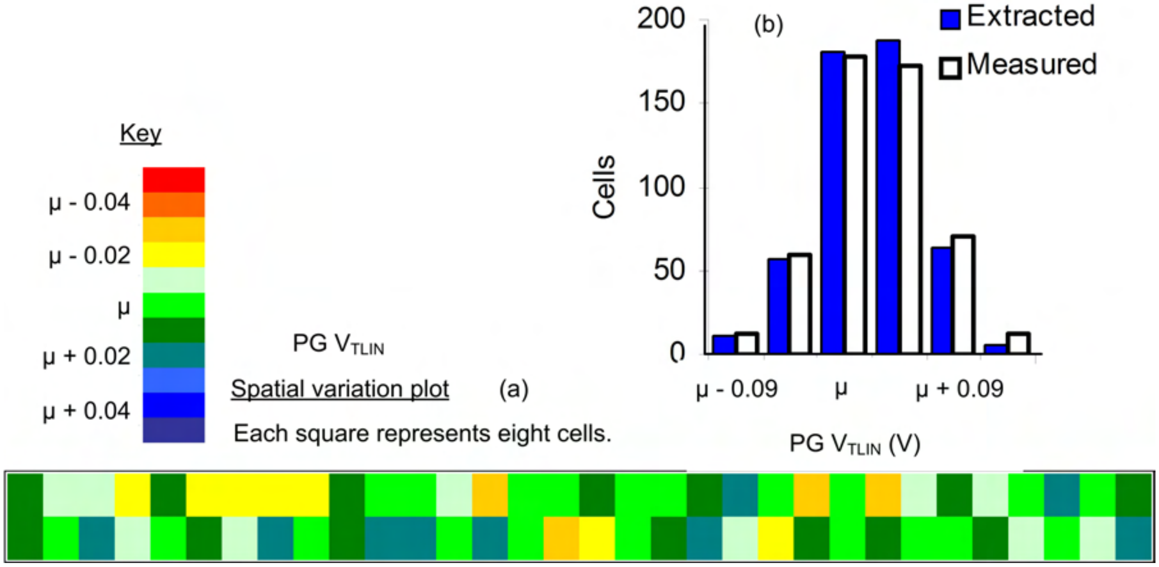

Figure 4.19. Measurements of **PG** V_{TLIN} are extracted from a dense array using SRAM voltage metrics. Low spatially-correlated variation is observed (a), and the overall distribution is close, but slightly tighter, than expected (b).

from simulation, since they could not be measured on this implementation of padded out cells.

There are several ways to reduce extraction error. Foremost among them is to increase the number of SRAM metrics with linearly independent sensitivities. In this context, linearly independent describes a set of sensitivities that cannot be reproduced as a linear combination of other sensitivities. For example, the sensitivities of SNM at $V_{DD} = 0.6$V and $V_{DD} = 1.3$V are linearly independent because the relative importance of threshold voltage increases with low V_{DD}; however, the sensitivities at a third voltage ($V_{DD} = 1.0$V) can be closely approximated with a combination of these two. There is little to be gained from measuring at a third voltage, since it is not linearly independent from the other two. This is illustrated in Fig. 4.20a. The extraction error is calculated as the length of the vector sum of the error, averaged among all cells, and normalized by the square root of the number of device parameters. In other words, for a cell with N device parameters $\vec{x}$ and extracted parameters $\vec{x_e}$, the extraction error ϵ is

$$\epsilon = \frac{||\vec{x} - \vec{x_e}||}{\sqrt{N}} \tag{4.4}$$

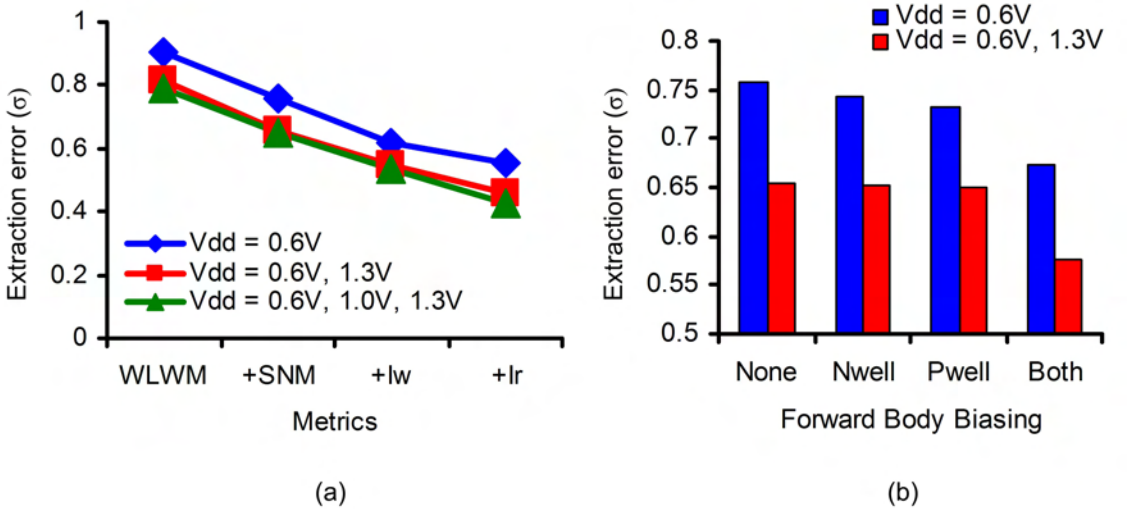

Figure 4.20. Extraction error can be reduced by using more linearly independent SRAM metrics or at least two different V_{DD} (a). It can also be improved by measuring with different well and substrate biasing; however, both n-well forward biasing and p-well forward biasing must be used to improve the fit.

With two V_{DD}, the extraction error is 0.1σ less than when only one V_{DD} is used. Adding a third V_{DD} results in only a marginal improvement. Extraction error is also reduced with every additional type of SRAM metric, such as SNM, I_W, or I_r. It is important to have both read and write metrics to improve the fits to both **PD** and **PU** devices. Each additional metric reduces the extraction error by approximately 0.13σ, and the effects are cumulative with multiple V_{DD}. Forward body biasing can also be used to increase the number of linearly independent metrics (Fig. 4.20b). Forward body bias (FBB) in the n-well strengthens the PMOS devices, increasing the sensitivities to **PU** parameters. FBB in the p-well strengthens the NMOS devices. Since the sensitivities are normalized, this has the effect of relatively lowering the sensitivities to the **PU** parameters. Individually, these effects make a small difference in the sensitivities, which is practically too small to distinguish from the variations caused by unextracted parameters. However, if both the metric with n-well FBB and the metric with p-well FBB are used, the extraction error can be lowered by approximately 0.08σ. Further reduction may be possible by varying temperature or frequency.

Extraction error can also be reduced with a couple of simple heuristics. Instead of extracting parameters from the least-squares fit, which minimizes $||A\vec{x} - b||$, the device parameters can be found by minimizing $||A\vec{x} - b|| + \exp(||\vec{x}||^2)$. The additional exponential term favors solutions with low amounts of variability, which are more likely to occur in practice, over solutions with unrealistically high standard deviations in a couple parameters that provide a better fit to the metrics. This heuristic is more useful for extracting device parameters to which the SRAM metrics have lower sensitivities. A second heuristic can be applied by scaling the SRAM metrics by a confidence factor. Some metrics like I_r are known to be well represented by the chosen device parameters, while others like SNM are better represented by other parameters. Multiplying the SRAM metric by a scaling factor (e.g. the inverse of the standard deviation in the error of Fig. 4.12) favors fits with high-confidence metrics. These heuristics can provide small improvements for extractions using a low number of SRAM metrics. In general, though, good accuracy can be obtained without these techniques by careful selection of the SRAM metrics.

This work demonstrates that process characterization is possible using only SRAM measurements. High data counts can be obtained quickly with a minimum of measurement sweeps. The extraction has good accuracy when both voltage and current metrics are used, and it is still reasonable if only voltage measurements are used. The accuracy can be further improved by increasing the number of independent metrics used in the extraction. It may be possible to modify BIST circuits to automate the measurements of these metrics. The statistical device data provided by such a circuit would be useful for process development and diagnosing failures without the costs of separate test runs or nanoprobing. It could also inform adaptive techniques in a manner akin to the variation sensor of Section 4.2, as a way of using feedback to improve SRAM robustness.

4.4 Independently gated FinFET SRAM designs

Besides adjusting the read / write tradeoff to correct for device variations, feedback can be used dynamically to improve the read and write margins based on the state of the cell. Implementation at the cell level requires new bitcell designs with different connectivities

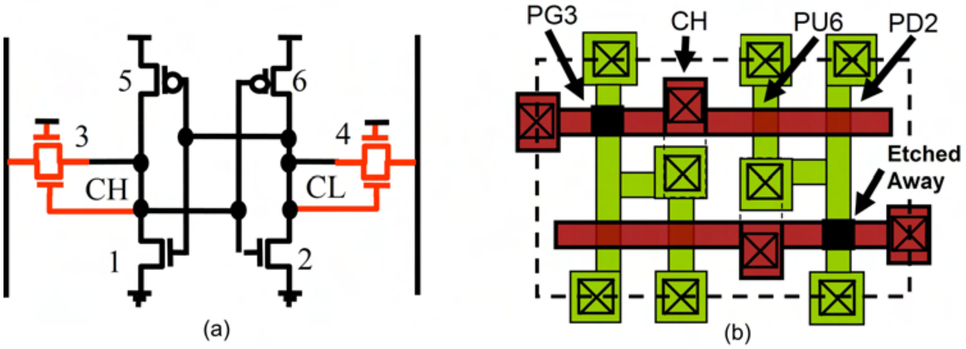

Figure 4.21. Using independently-gated FinFETs, a 6-T SRAM cell can be designed with the back gates of the **PG** transistors connected to the storage node, for enhanced read margin (a). The design can be implemented without area penalty, by simply extending the gates of the inverters to the appropriate **PG** (b). [6]

than a regular 6-T SRAM. For single gate devices or (devices with multiple but electrically-connected gates), extra transistors can be used but incur an area penalty. With thin-body SOI transistors, the device strength can be modulated capacitively, for example to dynamically enhance SNM by strengthening the **PD** device [14]. In this implementation, the area penalty is limited only to well contacts.

In Chapter 3, an argument was made for new device architectures to improve SRAM robustness. In devices with undoped channels, V_{T0} variations due to dopants are greatly reduced, all but eliminating one of the leading causes of SRAM variability. One of the unique features of the FinFET device structure is the capability for independent gating. Independently gated FinFETs have been demonstrated by several different researchers. With two gate connections available on each transistor, new bitcell designs can be implemented that enhance read stability and write-ability simultaneously.

4.4.1 Pass-gate feedback

Fig. 4.21 illustrates the operation of a 6-T SRAM cell with pass-gate feedback (PGFB) using independently-gated FinFETs [6]. The storage nodes are connected to the back gates of the **PG** device, allowing the strength of the **PG** transistor to be selectively decreased. When the storage node is at a logical zero, the back gate of its **PG** is biased at 0V, decreasing its strength. This effectively increases the beta ratio during a read operation, allowing the **PD** device to keep the storage node at a lower voltage, and thus improving the read margin. During a write operation, with the stored bit a logical one, the back-gate connection initially helps the **PG** device discharge the storage node but turns off as the cell state flips. By extending the gates of the inverters to the **PG** devices, the back-gate connection can be made with no area penalty over the conventional DG 6-T SRAM cell design.

In conventional double-gate SRAM designs, the gate work function determines threshold voltages and thereby tradeoff the read and write margins. Because of the processing challenges associated with work function tuning, it is expected that a single gate material with a single work function will be used for both NMOS and PMOS. The work function will be near the midgap of silicon, with higher work functions strengthening the PMOS devices and weakening the NMOS devices. This change improves read stability by increasing the trip point of the inverter, but it doubly decreases write-ability by weakening the **PG** and strengthening the **PU** device.

It can be shown that the PGFB SRAM design offers a more favorable read/write tradeoff than work function tuning [15, 16]. PGFB enables a higher SNM for read at high V_{DD} than is achievable with work function tuning alone. Furthermore, write-ability is enhanced for a matched SNM. In addition to enhancing the nominal margins, the PGFB design exhibits reduced sensitivities to process variations, which results in higher-yielding cells. Fig. 4.22 illustrates nominal SNM over a range of V_{DD} for a conventional double-gate 6-T SRAM cell and a cell with PGFB. The data in Fig. 4.22 is generated by simulation using the SRAM model of Chapter 2 and the Taurus device simulator to generate the FinFET I-V targets [17]. Device parameters are provided in Table 4.2. Mixed-mode Taurus simulations are also performed for validation and show good agreement with those of the SRAM model.

Gate work function tuning is used to make the conventional design as stable as the PGFB design at V_{DD} = 0.7V. To match the large SNM enhancement of PGFB, a higher work function (Φ_m = 4.82eV) is required for the conventional design. Although the two designs have comparable SNMs up to 0.7V, at higher V_{DD} work function tuning is less

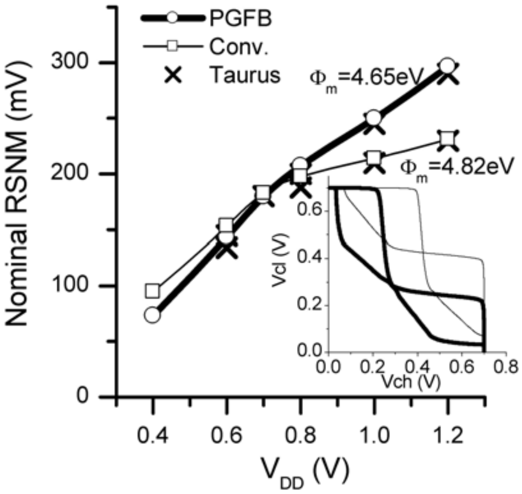

Figure 4.22. Even though gate work function tuning is used to match Read SNM at $V_{DD} = 0.7$V, the PGFB design has much higher read stability at high V_{DD}.

Table 4.2. FinFET Device Parameters

Parameters	FinFET	Bulk
L_G (nm)	22	22
Spacer length (nm)	24	24
T_{ox} (Å)	11	11
T_{Si} (nm)	15	N/A
Channel Doping (cm^{-3})	10^{16}	4×10^{18}
H_{FIN} (nm)	30	N/A
S/D doping gradient (nm/dec)	4	4

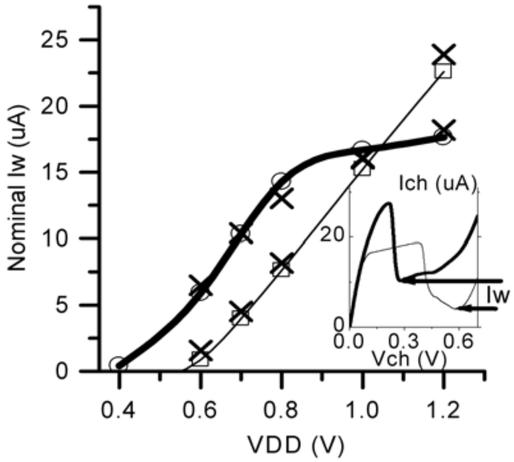

Figure 4.23. Nominal I_W for a conventional double gate FinFET SRAM cell and a PGFB cell with Φ_m values as in Fig. 4.22. The improvement in PGFB I_W at low V_{DD} is largely attributable to the lower gate work function; however, at higher V_{DD}, the feedback limits the **PG** current and prevents further I_W increase.

effective than PGFB. This is because the increasing effects of DIBL at higher V_{DD} values lower the gain of the inverter and reduce the benefit to SNM. For $V_{DD} > 1$V, the PGFB design achieves very high SNM, approaching 300mV, whereas the conventional design is limited to 230mV.

Though the effect of increasing the gate work function is limited for high supply voltages and introduces process complexity, its largest drawback is in the tradeoff with write-ability. Fig. 4.23 illustrates the write-abilities of the conventional double-gate and PGFB SRAM designs with matched SNM at 0.7V. At low V_{DD}, the conventional design exhibits reduced write-ability because its higher work function raises the NMOS V_{T0} and keeps the **PG** device in subthreshold operation. The PGFB design, with its inherently improved read stability, enables a better read/write tradeoff by allowing for a lower gate work function and therefore higher I_W. At high V_{DD}, the write-ability comparison is more complex. Nominal write-ability for the PGFB design saturates as the effect of the lower gate work function becomes less significant than the reduced drive on the back-gate. The bias on the back-gate, indicated by V_{CH} at the I_W point in the inset of Fig. 4.23, is lowered in the PGFB design and can be approximated as $V_{DD}/2$. This is an indirect effect of the lower gate work function, which reduces the trip point of the inverter, but it has an interesting benefit to the cell yield.

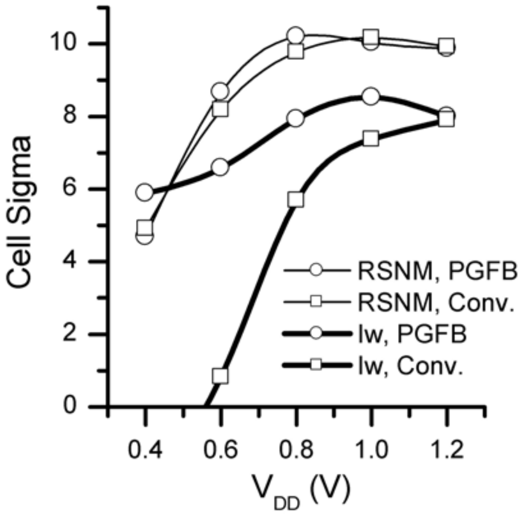

Figure 4.24. Projected yield considering read SNM and I_W independently. The large enhancement of PGFB I_W at low V_{DD} enables six sigma yield at 0.5V.

The projected cell yield in the presence of statistical variations for all six devices is illustrated in Fig. 4.24. L_G and T_{Si}, the thickness of the fin, are considered with $\sigma_L = \sigma_T = 1.54$nm. Because an undoped channel is used, dopant-induced variations to V_{T0} are neglected. With matched SNM at 0.7V, the conventional double-gate and the PGFB designs show comparable read cell sigmas across all values of V_{DD}. The yield saturates at ten standard deviations, which corresponds to the fin thickness. The PGFB design shows a significantly better write cell sigma at low V_{DD}. Much of this benefit is due to the improved nominal write-ability current observed in Fig. 4.23; however, the write yield also exhibits a low sensitivity to V_{DD}. This enables a wider range of operating voltages for the cell in an array and is in contrast to double-gate FinFET and bulk-Si MOSFET SRAM designs, which are often write-limited at low supply voltages. The reason for the low sensitivity is the reduced bias on the back gate at the write-ability point. Whereas the **PG** device in the conventional cell sees a variation in V_{DD} on both its front and back gates, in the PGFB cell it sees only half the variation on its back gate, which reduces the sensitivity to V_{DD}. This is an easily overlooked tradeoff of conventional gate work function tuning: increasing Φ_m not only degrades write-ability, but it also degrades the yield faster at low V_{DD}. Although

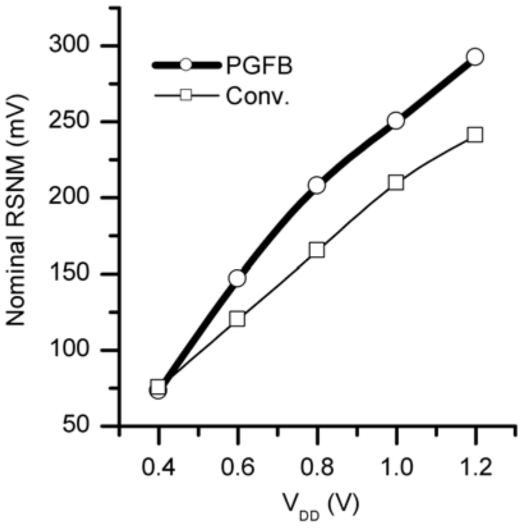

Figure 4.25. Nominal read SNM with gate work function tuning to match I_W at each V_{DD}. The PGFB design has approximately 20% greater read SNM for most values of V_{DD}.

PGFB enables a low gate work function, if a higher work function were used (perhaps for further read enhancement) the same degradation would be seen at low V_{DD}.

The read/write tradeoff of PGFB can be further explored by comparing SNM at matched write-ability. In Figure 4.25, work function tuning was used on the conventional double-gate design to match the PGFB write-ability at each point over a large V_{DD} range. For $V_{DD} > 0.4$V, the PGFB SNM is consistently higher by approximately 20%.

The biggest drawback of PGFB is that the weakened **PG** can degrade read performance. In this work, read performance is measured by the DC read current, that is, the current flowing through the **PG** transistor on the logical zero side of the cell when the **WL** and **BL** voltages are high. In some sense, this is a fundamental tradeoff; the reduced current that degrades read performance also decreases the charge that helps destabilize the cell. Fortunately, the degradation is not severe, especially when compared to a conventional design with matched SNM up to $V_{DD} = 0.8$V (Fig. 4.26). Although the PGFB design has only a single gate inverting the channel, the lower work function reduces V_T and increases the drive current. Furthermore, the zero storage node in the PGFB design stays closer to **GND** than in the conventional double gate design (Fig. 4.22), thus giving the back-gated **PG**

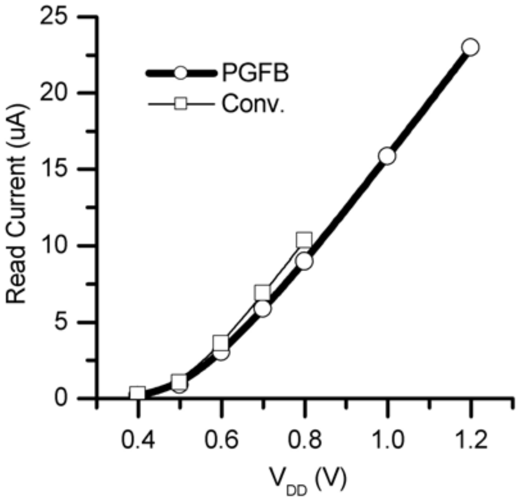

Figure 4.26. DC read currents with Φ_m chosen such that RSNM is matched at each V_{DD}. The PGFB design has $\approx$ 15% less read current than the conventional design at low V_{DD}. Above 0.8V, the conventional design cannot match PGFB SNM with any amount of gate work function tuning.

transistors more gate overdrive. The net result is only a 15% degradation in read current. It should be noted that for $V_{DD} > 0.8$V, the conventional design was unable to match the PGFB SNM with any amount of work function tuning.

The PGFB design therefore offers an improved read / write tradeoff versus work function tuning or threshold voltage adjustment. It provides higher SNM at equal write-ability or vice-versa, and it is therefore expected to be more robust to variations. Furthermore, it can be implemented without area penalty and a modest decrease in read access speed. With PGFB, work function tuning can still be used to enhance write-ability and optimize the read and write yields.

4.4.2 Pull-up write gating

A better approach to enhance write-ability is to selectively weaken the **PU** devices. Just as feedback can be used to weaken the **PG** transistor during a read operation, it is possible to increase write-ability by weakening the **PU** transistors during a write operation [15, 16]. This can be achieved using independently gated FinFETs for the PU devices and connecting their back gates to a write word line (**WWL**) (Fig. 4.27). During a write operation, setting

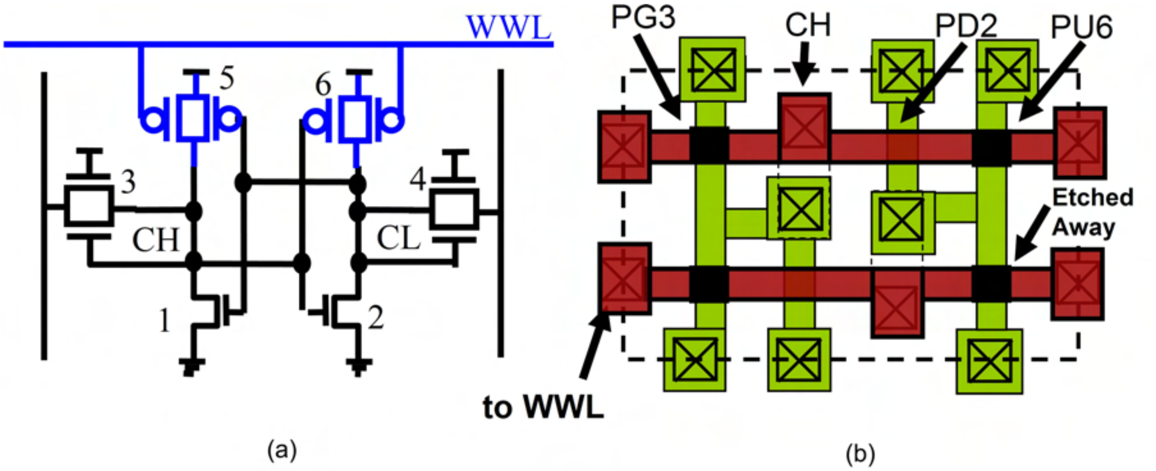

Figure 4.27. Schematic (a) and layout (b) for a 6-T FinFET SRAM with both PGFB and PUWG. An additional **WWL** contact is added at the edge of the cell, but does not consume any additional area.

$V_{WWL} = V_{DD}$ reverse-biases the back-gate of the **PU** devices, weakening them and hence increasing the write-ability current. At all other times, V_{WWL} is set to an intermediate value ($0 < V_{WWL} < V_{DD}$) to enable large SNM and hold margins. Both PUWG and PGFB can be implemented simultaneously with no cell area penalty as illustrated in Figure 4.27b. The **WWL** contacts are located next to the word line contacts and are shared between adjacent cells. The **WWL** is routed horizontally, but must interleave with the **WL** to make all the contacts. This requires an extra metal layer, but the cell area is not increased.

The PUWG design provides for enhanced write-ability, as shown in Fig. 4.28. The gate work function is chosen to provide 180mV RSNM at V_{DD}=0.7V. The PUWG design has a greater write-ability than the conventional DG design at all supply voltages, by as much as 60% at $V_{DD} = 1.2$V. With PGFB, a lower work function can be used to further improve write-ability at low V_{DD}. Whereas the PGFB design alone saturates in I_W at high V_{DD}, together with PUWG it achieves higher write-abilities with increasing V_{DD}.

During a read operation, the **WWL** voltage is lowered to an intermediate bias value. The choice of this bias value affects the voltage transfer characteristics of the cell. Fig. 4.29 shows the impact of the **WWL** bias on the voltage transfer curves. Setting $V_{WWL} = V_{DD}$ weakens the **PU** device, thereby lowering the trip point of the inverter. On the other hand,

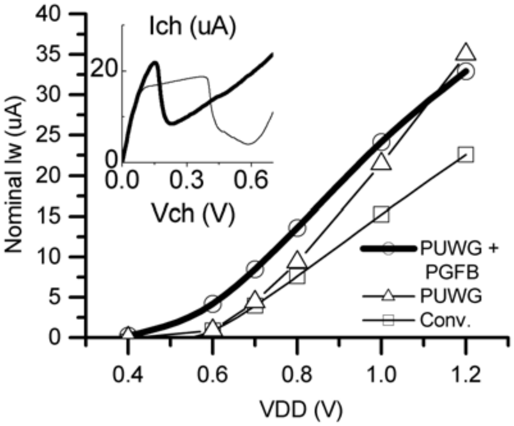

Figure 4.28. The PUWG design enhances write-ability for all V_{DD} ($\Phi_m = 4.87$eV), while the addition of PGFB enables a lower $\Phi_m = 4.67$eV for further improvement at very low V_{DD}.

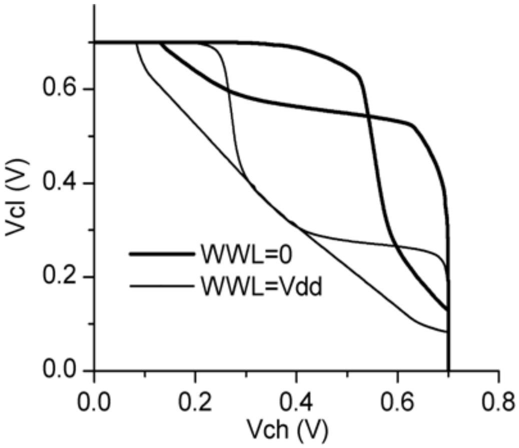

Figure 4.29. When **WWL** is low, the **PU** leaks current and the top shoulders of the curves are pulled out. This effect can complement that of PGFB in keeping the internal node voltage closer to ground.

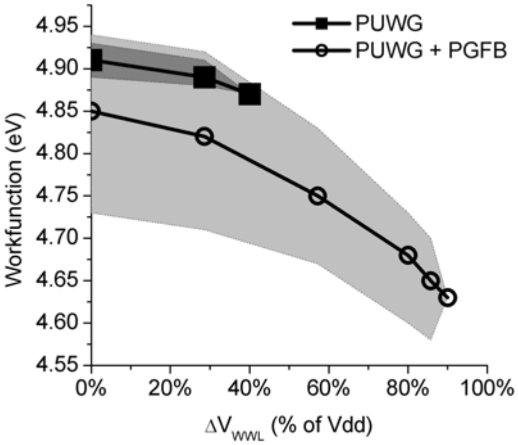

Figure 4.30. Peak and range (shaded) of gate work function Φ_m as a function of V_{WWL} such that RSNM > 180 mV at $V_{DD} = 0.7$V. Above 85%V_{DD}, the **WWL** gate turns on the **PU** devices, sharply degrading the SNM. With PGFB, a much lower Φ_m is needed for RSNM > 180mV, thereby improving cell write-ability.

setting $V_{WWL} = 0$V fully turns on the back-gate of the **PU** device, thereby pushing out the upper shoulder of the voltage transfer curve (increasing the maximum V_{CH} for $V_{CL} \approx V_{DD}$). The effect complements that of PGFB in increasing the SNM of the cell: while PGFB boosts SNM by lowering the node settling voltage during a read, **WWL** biasing can increase the trip point of the inverter and achieve an increase in the SNM as well. The largest SNM is obtained when the trip point of the inverter is close to $V_{DD}/2$. There is an optimal **WWL** bias that maximizes SNM near this point.

The bias value to which **WWL** steps down after a write operation determines the range of work functions that meets the given SNM target (Fig. 4.30). A larger step ΔV_{WWL}, measured as a fraction from V_{DD}, corresponds to a lower voltage on the PMOS back-gate and an increased trip point of the inverter. A larger ΔV_{WWL} enables a lower work function to achieve the same SNM. At 0.7V, the range of values for SNM > 180mV is limited. With PGFB, the range is much larger, enabling low work functions even at moderate biases. Even with PGFB, though, a large enough ΔV_{WWL} can cause the back-gate to dominate the PMOS current and diminish SNM. Therefore in order to minimize the impact on SNM and maintain high yield at low V_{DD}, ΔV_{WWL} should be kept to a moderate value.

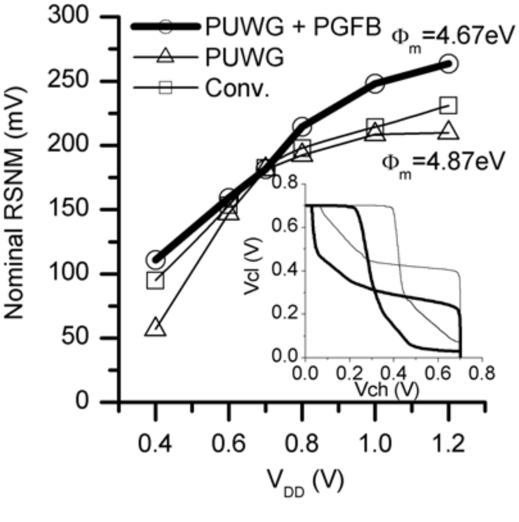

Figure 4.31. Nominal read stability with gate work functions chosen such that RSNM = 180mV at $V_{DD} = 0.7$V. At high V_{DD} the effect of PGFB on the butterfly curves complements that of PUWG (inset), resulting in greater RSNM.

Fig. 4.31 illustrates high nominal SNM for the combination of PGFB + PUWG with $V_{WWL} = 0.4$V during a read operation. At very low and high V_{DD}, the SNM is highest for the combination of PGFB + PUWG due to their complementary effects on the butterfly curves. The inset of Fig. 4.31 compares butterfly curves at $V_{DD} = 0.7$V for the combination (bold curves) with the conventional design (thin curves). The inverter trip point for PGFB + PUWG is closer to $V_{DD}/2$, due to a lower gate work function and the effects of PUWG. The lower shoulders of the butterfly curves exhibit less linearization from the PG, due to PGFB. The slope of the curves near the trip point is somewhat degraded due to the reduced PMOS gain of PUWG; however, the effect on SNM is small.

A high nominal SNM increases the read yield for the PGFB+PUWG combination, particularly at low V_{DD} (Fig. 4.32a). The read yields for all designs are comparable, but is slightly lower for the PUWG-only design at low V_{DD}. The PMOS transistors are partially on for this design during the read, so they are more sensitive to device parameter variations than in other designs. The high nominal SNM achieved with the addition of PGFB offsets this effect and enables yield of over six sigma at $V_{DD} = 0.4$V. The I_W yield is highest for the PGFB + PUWG design as well, primarily due to the enhancement in nominal write-ability

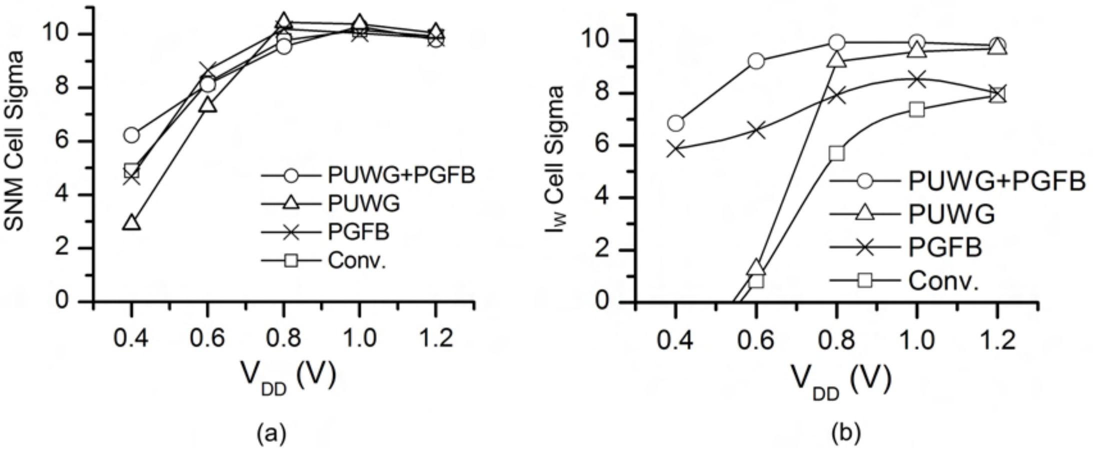

Figure 4.32. Projected yield for read stability (a) and write-ability (b). The designs are comparable in SNM cell sigma at higher V_{DD}, but the combination of PGFB and PUWG provides for higher yield at lower V_{DD}. The combination also achieves greater than six sigma yield at very low voltages, due to a lower gate work function.

(Fig. 4.32b). In the PUWG design, since the PMOS transistors are weaker during the write, they are less sensitive to process variations and yield is further enhanced. With the low work function enabled by PGFB, yields greater than six sigma are achievable down to 0.4V.

Independently gated FinFETs are therefore attractive for SRAM not only for the reduced V_{T0} variation of a multi-gate undoped device architecture, but also for the new bitcell designs it enables. Designs such as PGFB and PUWG can be used to simultaneously improve read and write yields by two sigma or more, without penalty to cell area. This enhancement is cumulative with many other device and circuit designs in this work and could be large enough to extend SRAM scaling for several technology generations.

4.5 Summary

This chapter presents three different closed-loop approaches to improving SRAM robustness. Each approach operates on a different scale. Variation sensors, which measure spatially correlated variations, can improve yield for blocks of cells up to a chip-wide level. Device characterization techniques, which collect large amounts of statistical data on process variations, can be useful for debugging or developing a process on the large scale of a batch of wafers. At the other extreme, independently-gated FinFETs tradeoff cell stability for write-ability depending on the logical state, for each individual cell.

Each of these approaches is designed to minimize area overhead. The variation sensors consist of resistive dividers which fit into bitcell layouts. The required number of cells is self-limiting in the sense that variability which is too small to be measured with a few sensors will be likewise too small to degrade yield. Additionally, the variation sensor requires an amplifier to generate the optimal $V_{WL}*$; however, the amplifier size can be amortized over multiple sensors on a chip or possibly reduced by using a discrete set of control voltages. The overhead could possibly be as small as an extra row in an array plus the circuit needed to generate the wordline signal.

By measuring the array bitcells themselves, the device characterization technique does not require the overhead of large device arrays in order to get high counts for meaningful statistics. Reasonable estimation of actual device I-V targets can be obtained from SRAM voltage metrics alone, which further reduces the area required for on-chip testing.

The least area overhead of all is obtained with independently-gated FinFETs. PGFB dynamically enhances read stability, with an improved read / write tradeoff. PUWG dynamically enhances write-ability to further improve yield. Both designs have no cell area penalty–instead relying on process innovation–but the PUWG design may require extra peripheral circuitry to drive its write wordline. A significant advantage of the FinFET SRAM designs is the expected yield improvement from using undoped channels, as discussed in chapter 3.

These approaches are all complementary, with each other as well as with many other techniques–circuit and device–for reducing SRAM variability. The independently-gated FinFET designs exemplify the potential for yield enhancement when device and circuit techniques are combined. Combining this approach with systematic variation sensors could provide further enhancement, particularly since FinFETs are expected to be more sensitive to lithography variations, a source of systematic variability. There are many promising combinations–for both the short term and the long term–and the optimal choice depends on external factors, such as the product's application and its market.

4.6 References

[1] D. Lammers and R. Wilson. Soft errors become hard truth for logic. *EE Times*, May 3, 2004.

[2] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr. A 153MB-SRAM design with dynamic stability enhancement and leakage reduction in 45nm high-κ metal-gate CMOS technology. *International Solid-State Circuits Conference*, pages 376–377, 2008.

[3] H. Pilo, J. Barwin, G. Braceras, C. Browning, S. Burns, J. Gabric, S. Lamphier, M. Miller, A. Roberts, and F. Towler. An SRAM design in 65nm and 45nm technology nodes featuring read and write-assist circuits to expand operating voltage. *IEEE Symposium on VLSI Circuits*, pages 15–16, 2006.

[4] L. Chang, Y. Nakamura, R.K. Montoye, J. Sawada, A.K. Martin, K. Kinoshita, F.H. Gebara, K.B. Agarwal, D.J. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek. A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS. *IEEE Symposium on VLSI Circuits*, pages 252–253, 2007.

[5] A. Kawasumi, T. Yabe, Y. Takeyama, O. Hirabayashi, K. Kushida, A. Tohata, T. Sasaki, A. Katayama, G. Fukano, Y. Fujimura, and N. Osuka. A single-power-supply 0.7V 1GHz 45nm SRAM with an asymmetrical unit-β-ratio memory cell. *International Solid State Circuits Conference*, page 21.4, 2008.

[6] Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. King, and B. Nikolic. FinFET-based SRAM design. *IEEE International Symposium on Low Power Electronics and Design*, pages 2–7, 2005.

[7] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, M. Igarashi, M. Takeuchi, H. Kawashima, H. Makino, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, K. Ishibashi, and H. Shinohara. A 65nm SoC embedded 6T-SRAM design for manufacturing with read and write cell stabilizing circuits. *Symposium on VLSI Circuits*, pages 17–18, 2006.

[8] K. Agarwal, F. Liu, C. McDowell, S. Nassir, K. Nowka, M. Palmer, D. Acharyya, and J. Plusquellic. A test structure for characterizing local device mismatches. *IEEE Symposium on VLSI Circuits*, pages 67–68, 2006.

[9] H. Yu, Y.-S. Kim, Y.-G. Kim, H.-C. Kim, U.-R. Cho, and H.-G. Byun. A SRAM core architecture with adaptive cell bias scheme. *IEEE Symposium on VLSI Circuits*, pages 128–129, 2006.

[10] Z. Guo, A. Carlson, L.-T. Pang, T.-J. King-Liu, and B. Nikolic. Large scale read/write margin measurement in 45nm CMOS SRAM arrays. *IEEE Symposium on VLSI Circuits*, 2008.

[11] X. Deng, W. K. Loh, B. Pious, T. W. Houston, L. Liu, B. Khan, D. Corum, J. Raval, J. Gertas, F.-Y. Rousey, J. Steck, C. Suwannakinthorn, and R. McKee. Characterization of bit transistors in a functional SRAM. *IEEE Symposium on VLSI Circuits*, 2008.

[12] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake. Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit. *International Solid State Circuits Conference*, page 34.5, 2006.

[13] D. Burnett. Statistical design issues of SRAM bitcells and sense amps. *IEEE Silicon on Insulator Conference*, 2006. Short Course.

[14] M. Yamaoka, K. Osada, R. Tsuchiya, M. Horiuchi, S. Kimura, and T. Kawahara. Low power SRAM menu for SOC application using yin-yang-feedback memory cell technology. *Symp. VLSI Circuits*, pages 288–289, 2004.

[15] A. Carlson, Z. Guo, S. Balasubramanian, L.-T. Pang, T.-J. King, and B. Nikolic. FinFET SRAM with enhanced read / write margins. *IEEE Silicon on Insulator Conference*, pages 105–106, 2006.

[16] A. Carlson, Z. Guo, S. Balasubramanian, R. Zlatanovici, T.-J. King Liu, and B. Nikolic. SRAM read / write margin enhancements using FinFETs. *Trans. Very Large Scale Integration*, page to be published, 2008.

[17] TAURUS is a trademark of Synopsys, Inc.

Chapter 5

Conclusion

Variations present a formidable challenge for SRAM scaling in future technology nodes for two reasons. First, the magnitude of device variations is increasing. Random dopant fluctuations, which are expected to remain the most significant component of threshold voltage variation until gate lengths scale below 20nm [1], follow a $1/\sqrt{WL}$ dependence. Line edge roughness and other lithography variations do not scale with line widths, but become relatively more significant. They are expected to dominate variability for sub-20nm gate lengths and multi-gate transistors. The second reason that makes variations such a formidable challenge is the exponential increase in memory size with each product generation. Microprocessor caches currently require more than 10^7 bitcells [2], a number so great that even exceptionally rare events can have a noticeable impact on product yield. Moreover, this number is increasing, as microprocessor designs switch to multi-core architectures with higher demands for memory.

The problem of reducing variation to enable continued SRAM scaling is complex, but solvable. First of all, a thorough understanding of the mechanisms by which device variations cause SRAM failure is essential. Fast and accurate modeling reveals an intricate web of tradeoffs: write-ability vs. read stability, stability vs. area, area vs. variability, and so on (Fig. 5.1). New transistor structures can be designed to improve these tradeoffs by

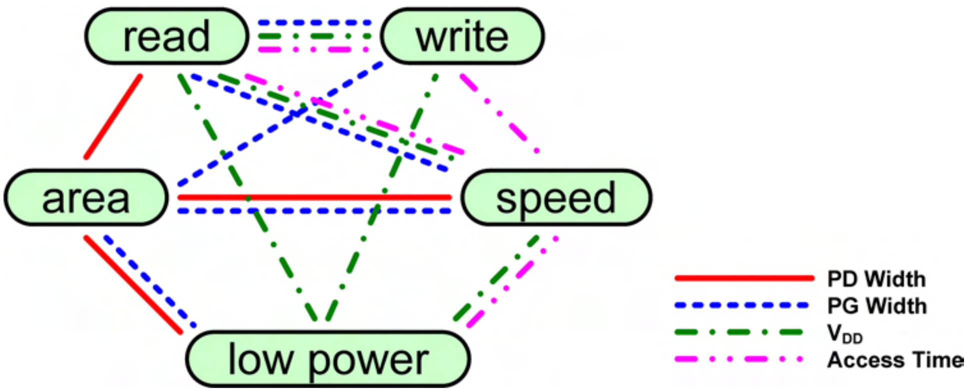

Figure 5.1. SRAM design is complicated by a web of tradeoffs, associated with changing **PD** or **PG** width, supply voltage, or timing.

reducing the sources of variation, for example by using undoped channels. Feedback circuits can be used to compensate for variability after fabrication with less cost and risk.

There are many possible solutions to enable continued SRAM scaling. This work establishes a modeling framework for evaluating them and suggests several promising alternatives. It is hoped that this work can inform future research and development of SRAM.

5.1 Contributions of this work

This work contributes specifically to three aspects of reducing variability in SRAM: an understanding of the link between process variations and SRAM failure, informed by analysis of recent new effects; device and process technologies that address the causes of transistor variation; and circuit designs employing feedback to enhance robustness.

SRAM variability can be traced to specific device parameters with sensitivity analyses. Recently, write margins have fallen to comparable levels with read margins, requiring a shift in focus from enhancing read stability to optimizing its tradeoff with write-ability. Of paramount concern is the overall yield for large arrays, which requires accurate modeling of extremely rare statistical events. Simple mean / sigma estimates rely too heavily on assumptions of normality and have been shown to be metric-dependent. An iterative algortihm based on SRAM simulations near conditions of failure provides a fast and robust estimate. Yield analyses with this algorithm can cover the major sources of process-induced transistor variation, as well as time-dependent sources of variation, such as NBTI. Statistical design methodologies should consider all of these sources.

DC yield simulations cannot replace transient analyses with characterized compact models using many parameters, but they can provide ballpark estimates to evaluate speculative device technologies. New device architectures and processes will probably be necessary to reduce sources of variation such as random dopant fluctuation and line edge roughness. Multi-gate devices, such as tri-gate bulk or FinFETs, can provide excellent short-channel control with undoped or very lightly doped channels, drastically reducing threshold voltage variation. Spacer-defined active and gate layers–or, indeed, entire SRAM arrays–can provide smaller, more robust cells than can be achieved with current photolithography technology.

Because the cost and risk of new device technology is significant, circuit-based solutions are expected to meet at least the short-term need for robustness. BIST circuits that automate the collection of large amounts of SRAM data can be expanded to collect voltage and current metrics under different bias conditions for each cell. Individual transistors can then be characterized from the cell measurements, providing high confidence distributions in device parameters over a large range of variation. Feedback circuits can be used to compensate for spatially correlated variations with optimal biasing. With independently gated FinFET devices, new SRAM bitcells can be designed to simultaneously improve read and write yields.

As exemplified by the FinFET SRAM, a combination of device and circuit techniques can be used, with complementary enhancements. Many techniques proposed or analyzed in this work have the potential to increase SRAM yield by a few sigma. With combinations, it is expected that SRAM scaling can be extended far into the future, possibly to the end of the roadmap.

5.2 Future directions

Although this work informs of promising directions for SRAM scaling, there is much work still to be done.

The DC SRAM model developed in this work improves accuracy because it does not assume Gaussian cell metrics; however, it does assume normally distributed device parameters. In fact, the tails of these distributions may not be Gaussian in all cases; however, a methodology using the SRAM as a process characterization vehicle can provide the true distributions. The mathematics of the model can possibly be extended to the case where device parameter distributions are not Gaussian but known. Additionally, the *cell sigma* metric is only a rough approximation of actual array yield. More rigorous mathematics might be able to relate cell sigma to a confidence interval for the true yield. The model can also be extended to handle both random and systematic variations easily if the respective distributions are known. Other time-dependent sources of reliability failures, such as time-dependent dielectric breakdown or positive bias temperature instability can also be considered.

Device technologies like triple-gate bulk transistors, independently-gated FinFETs, and spacer lithography are expected to reduce variation at its sources, but they have not yet been shown experimentally to improve SRAM yield at large integration scales. There are several process and design challenges associated with each technology that can introduce significant additional variations before they are mastered. The additional variability can mask the benefits of the new technology until the technology matures. Due to the need for sub-50nm dimensions to observe reductions in random dopant fluctuation and line edge roughness, such technologies are more efficiently developed in an industrial setting.

The variation sensor circuit in Section 4.2 can also be enhanced. Instead of controlling for orientation-dependent variation, the sensor can be designed with sensitivity to these effects. Compensation can be made for these effects using row- or data-dependent biasing.

In addition to wordline biasing, several other biasing or timing approaches could be used to compensate for variations.

5.3 Final Thoughts

SRAM scaling is the current focus of much research, and the above issues will undoubtedly be addressed. This work presents several viable options to enable SRAM scaling to continue, but some are more likely to be implemented than others. The candid perspective offered below is informed by this work and current trends, but by no means are the predictions guaranteed.

There is some question whether SRAM scaling will be pursued at all, or whether alternative memories, such as floating body RAM or embedded DRAM will replace it. Although these technologies may find a limited place among the levels of microprocessor cache, they will find it difficult to equal the speed and relative stability of SRAM. Floating body RAMs are currently notoriously slow; even if their speed can be increased, they are feared to be even more vulnerable to process variations. Embedded DRAM is fast, but difficult and expensive to manufacture. SRAMs are therefore likely to remain a critical component of microprocessors.

In the short term, incremental solutions will probably be preferred to increase SRAM robustness. SRAM designs will increasingly feature bias or timing circuits to improve the read / write tradeoff. Dynamic noise margins, although more difficult to model, will be preferred for their more realistic estimations of yield. SRAM layouts will likely become more and more regular, as in the 45nm cell of [3]. Currently multiple photolithography steps are used to define, then cut, gates to reduce rounding; in the future, this practice could extend to the active layer as well. The optimization of high-κ dielectrics will provide a short respite from increasing threshold voltage variation.

Ultimately, though, these techniques will not be able to compete with the $1/\sqrt{WL}$ behavior of threshold voltage scaling. Multi-gate architectures will be introduced with very low channel doping. Threshold voltages will be set by gate work functions, leveraging the research from high-κ gate stacks. Triple-gate bulk transistors are probably more likely to be implemented first, due to their easier manufacturability compared to FinFETs. If necessary, the triple-gate bulk design may evolve toward that of a FinFET. Independently-gated FinFETs will probably be considered too difficult and risky to manufacture on a large scale; however, they may possibly be implemented in lieu of gate length scaling if FinFET processes are well-established.

The major caveat in these predictions is economics. The cost of technology scaling grows with each generation, as more advanced tools are required. A transition to a multi-gate device technology is expected to be particularly expensive, with disruptions to both process development and design. In addition, recent developments in multi-core architectures could outpace the ability of software to exploit their advantages. While increasing numbers of logic cores may maintain demand in the profitable server market, the lack of benefits in the personal computer and rapidly expanding ultra-portable markets may forestall continued scaling. Although there is intriguing long-term potential for massively parallel computing systems, it is likely very far off. Correspondingly, the technology scaling rate may slow down to allow microprocessor development to focus on new architectures or systems.

The scaling rate notwithstanding, memory demand is likely to steadily increase for the foreseeable future. Continued SRAM development will be important to ensure high yields and good performance. The device and circuit techniques investigated here, individually or in combination, can direct that development to enable SRAM scaling for the next several product generations.

5.4 References

[1] A. Asenov. Simulation of statistical variability in nanoscale MOSFETs. *VLSI Tech. Dig.*, pages 86–87, 2007.

[2] Specifications for the Core 2 Duo. Intel.

[3] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr. A 153MB-SRAM design with dynamic stability enhancement and leakage reduction in

45nm high-κ metal-gate CMOS technology. *International Solid-State Circuits Conference*, pages 376–377, 2008.

Index