Interactions of Technology and Design in Nanoscale SRAM

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

In partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

by

Randy W. Mann

December 2010

Approval Sheet

The dissertation is submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering

_____
Randy W. Mann (AUTHOR)

This dissertation has been read and approved by the examining committee

_____
Dr. Benton Calhoun (Advisor)

_____
Dr. Travis Blalock (Committee chair)

_____
Dr. Sudhanva Gurumurthi (Committee member)

_____
Dr. Mircea Stan (Committee member)

_____
Dr. Scott Barker (Committee member)

Accepted for the School of Engineering and Applied Science:

_____
Dean, School of Engineering
and Applied Science

(December, 2010)

Quotation

“I look to the diffusion of light and education as the resource to be relied on for ameliorating the condition, promoting the virtue, and advancing the happiness of man.”

- Thomas Jefferson (1822)

Research advisor
Benton H. Calhoun

Author
Randy W. Mann

Interaction of Design and Technology in Nanoscale SRAM

Abstract

Continued advances in silicon technology have enabled the VLSI industry to shrink the area of the transistor by roughly a factor of two with each successive technology node. This trend has continued unabated for the past five decades and has made personal computing devices ubiquitous in modern culture. Made possible by continuous advances in CMOS technology and fueled by a growing and fiercely competitive market, in order for this trend to continue, continued advances in CMOS process technology as well as circuit design innovation are required.

Reduced device dimensions and operating voltages that accompany technology scaling have led to increased design challenges with each successive technology node. Thus, reduced functional yield margins coupled with increasing variability of the CMOS device characteristics have become the most significant problem facing future nanoscale SRAM, motivating this effort.

To address these challenges, a custom scaled (90nm-22nm) predictive technology model

(PTM) based framework is developed, using published industry target values to quantify and address the challenges confronting nanoscale SRAM below the 65nm node. In addition to random variation sources, the role of scaling, use of pushed ground rules for bit cell design, and the 6T cell layout topology can contribute to non-random or systematic device mismatch. These sources of variation and the underlying mechanisms are examined using technology computer aided design (TCAD) tools and hardware measurements.

The 6T SRAM cell design has been successfully scaled in both bulk and silicon on insulator (SOI) technologies down to the 32/28nm node and has remained for more than a decade the dominant technology development vehicle for advanced CMOS technologies. While the industry has converged on a specific layout topology, which remains dominant in the VLSI industry, continued scaling may stimulate further investigation of alternate bit cell topologies. Based on an examination of the layout topologies used for the 6T bit cell, sources of systematic mismatch, and changing lithography constraints, a new topology for 6T SRAM beyond the 22nm node is proposed in this work.

While circuit assist methods have shown promise in extending the life of the 6T SRAM, this work develops a sensitivity based methodology for assessing the effectiveness of the assist methods in addressing the reduced functional margins. Additionally, a new margin/delay analysis is developed as a means of assessing the functional effectiveness of the circuit assist methods. The margin/delay analysis may be further extended to assess the limits of circuit assist methods in extending the 6T SRAM beyond 32nm node. Finally, a constraint based analysis is used to assess the extent to which these methods may provide effective solutions as the technologies are scaled beyond the 22nm node.

Acknowledgments

It has been a privilege to work with my Ph.D. advisor, Prof. Benton Calhoun, whose work in SRAM and sub-Vt CMOS inspired me to return to graduate school and pursue a doctorate in electrical engineering. Throughout the course of my graduate research at the University of Virginia, I sought out and highly valued his guidance, counsel and insights. I am grateful for his guidance through the entire process.

It is my pleasure to thank the members of my thesis committee, Prof. Travis Blalock, Prof. Mircea Stan, Prof. Scott Barker, Prof. Sudhanva Gurumurthi and Prof. Benton Calhoun for their time, encouragement through the process and for many helpful discussions, and suggestions.

I had the most fortunate experience to work closely with Dr. Jiajing Wang, and Satya Nalam, in the RLPVLSI SRAM group and benefited greatly from their knowledge, insights and interactions. Thank you for your many contributions to this work.

I also want to express my appreciation for my colleagues in the RLPVLSI group; Yousef Shakhsheer, Kyle Craig, Sudhanshu Khanna, Yanging Zhang, Joseph Ryan, Taeyoung Kim, James Boley, Alicia Klinefelter, and Aatmesh Shrivastava. Additionally I want to thank Nashant George, Stuart Wooters, Jerry Qi, Adam Cabe, Saad Arrabi, Jiawei Huang and Jonathan Bolus. This effort would have been unimaginably more difficult without the valuable discussions and assistance from this talented group.

I would like to specifically acknowledge my colleagues from industry and academia who have contributed in some way in this work: Joseph Wang, Terry Hook, Phung Nguyen, Geordie Braceras and Harold Pilo, Martin Ostermayr, Avik Ghosh, Aziz Bhavnagarwala, and Jeff Johnson for contributions and helpful discussions. I would also like to thank Jim Ryan, Rick Amos, Victor Ku, Gary Bronner and Subramanian Iyer for their helpful

discussions over the past several years.

I deeply appreciate the encouragement and patience of my family during this phase of my life; my parents, Chuck and Jean, my four children (Jonathan and his wife Ruolin, Timothy and his wife Laura, Cara and Christopher) and my brother, Larry and his wife Teresa. You have all played a large role in making this endeavor possible.

Finally, I would like thank my wife, Bonnie, for her support and encouragement throughout our journey through life. Her confidence in me, love and encouragement have been a constant in my life.

Contents

List of Figures

List of Tables

List of Acronyms

ALC Assist limit (margin/delay) contour

Avt Slope of device mismatch sigma values

BL Bit line

BLB Bit line bar

CDF Cumulative distribution function

CMOS Complimentary Metal-Oxide-Semiconductor

CMP Chemical mechanical polish

DFM Design for manufacturability

DIBL Drain induced barrier lowering

DPF Device packing factor

ECC Error correction circuitry

EUV Extreme ultraviolet

eV electron volts

FDSOI Fully depleted SOI

GWF Gate work function

HCI Hot carrier injection

GIDL Gate induced drain leakage

HSNM Hold static noise margin

LER Line edge roughness

LP Low power

MC Monte Carlo

MeV Mega electron volts

MOSFET Metal-Oxide-Semiconductor Field Effect Transistor

MOS Metal oxide semiconductor

MUGFET Multi gate field effect transistor

NBTI Negative bias temperature instability

NFET N-type field effect transistor

NMOS N-type metal oxide semiconductor

NWELL N-type doped region

OPC Optical proximity correction

PBTI Positive bias temperature instability

PDF Probability distribution function

PFET P-type field effect transistor

PMOS P-type metal oxide semiconductor

PTM Predictive technology model

PVT Process-temperature-voltage

PWELL P-type doped region

RDF Random dopant fluctuation

RV Random variable

RSNM Read static noise margin

SCE Short channel effect

SER Soft error rate

SEU Single event upset

SIT Sidewall image transfer

SNM Static noise margin

SOI Silicon on insulator

SRAF Sub-resolution assist feature

SRAM Static Random Access Memory

STI Shallow trench isolation

TCAD Technology computer aided design

Tox Gate oxide thickness

UTSOI Ultra thin SOI

Vt Threshold voltage

WL Word Line

WM Write margin

Vdroop Voltage droop

Vfwd forward voltage bias

Vgs Gate minus source voltage

VLSI Very large scale integration

Vmax maximum voltage tolerable as specified by the technology provider

Vmin Minimum voltage at which a given array of bits can successfully be written and read at a specified yield target

Vsb Source minus body voltage

Vtmm mismatch in threshold voltage

Chapter 1

Introduction: Technology Scaling and SRAM

1.1 Background and Motivation

The commercial success and and widespread accessibility of multiple computing platforms available today ranging from hand-held and portable devices to mainframe supercomputers has been made possible by the reduced cost per memory bit and logic gate with each successive technology generation. This reduced cost is made possible by continued advances in CMOS device scaling. The design challenges such as increased variability and quiescent power coupled with reduced noise margins are inherently linked to the industry scaling methodology. These challenges are even more pronounced in the dense SRAM devices which commonly employ sub-minimum design rules. SRAM remains the most cost effective embedded memory solution for many applications; however, fundamental challenges arise as technologies continue to scale below 100nm. This chapter defines the

rapidly emerging challenges facing CMOS SRAM technologies in the nanoscale era and defines the problem set to be addressed and the scope of the work.

The 6T SRAM cell design has been successfully scaled across many technology generations and, because it generally requires little deviation from base logic processing, is frequently used as the technology development vehicle for advanced CMOS technologies. For example, as we continue to scale beyond the 90nm node, the memory designer must account for significant increases in leakage mechanisms such as gate tunneling and gate induced drain leakage (GIDL) that were much less significant in prior nodes.

Despite these challenges, the 6T SRAM is expected to continue to play a dominant role in future technology generations because of its combination of density, performance, and compatibility with logic processing. The successful commercial scaling of the 6T SRAM driven by strong industry competition has followed a well defined linear shrink factor of $0.7\times$ over multiple generations resulting in a predictable $2\times$ reduction in cell area per generation. Despite numerous technical challenges in lithography, device, and process integration, the trend in 6T bit cell area is expected to continue beyond the 28/32nm node. This trend in 6T cell area, shown in Fig. 1.1, is projected beyond the 22nm generation. For example, the competitive 6T cell size is expected to be approximately $0.031\mu\text{m}^2$ at the 15nm node. This continued trend in area reduction is accompanied by the well known consequence of increased variability associated with the reduced channel area. Although technology options such as high-κ with metal gate have provided some relief in variability, the reduced functional margins and increased variation beyond the 28/32nm generation will drive further design and process technology innovation.

To investigate the impact of scaling in future technologies, predictive technology mod-

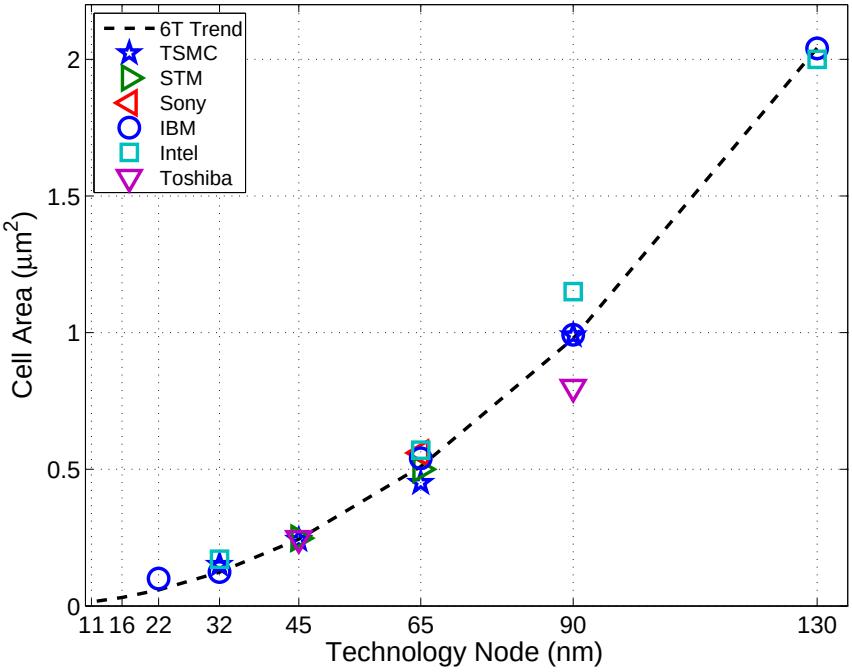


Figure 1.1: Trend in SRAM cell size with scaling based on published cell sizes.

els (PTMs) were customized to reflect the low power technology offerings available today for nodes from 90nm to 22nm. The models were calibrated based on published industry LP CMOS data [87] [55] [54] [88]. Fig. 1.2 provides the most critical metrics used and technology scaling assumptions for the LP models in our study.

1.1.1 Increasing Device Variation

Variation is both a well known limiter to scaling and fundamentally dependent on the specific process technology. Threshold voltage (Vt) variation due to random dopant fluctuations (RDF) in the device channel has been empirically shown to be proportional to $1/\sqrt{WL}$ as described by Pelgrom et al., [69]. Because of the use of narrow devices in the SRAM cell environment, the variation associated with RDF is a dominant variation

	Node	65		45		32		22	
	device type	N	P	N	P	N	P	N	P
Target Values	Vnom (V)	1.2		1.1		1.1		1	
	Tox (nm)	2		1.8		1.6		1.4	
	Lpoly (nm)	56		39		27		19	
	Ion (uA/um)	600	300	620	300	700	380	720	380
	Ioff (pA/um)	250		400		1000		2000	
	HVT Ion (uA/um)	400	210	410	210	440	340	450	340
	HVT Ioff (pA/um)	10		30		50		150	
Tuned model values	Ion (uA/um)	606	305	615	309	709	381	725	385
	Ioff (pA/um)	250	219	477	409	947	965	1858	1915
	HVT Ion (uA/um)	409	220	425	229	444	330	469	331
	HVT Ioff (pA/um)	10	9	36	35	45	62	183	172

Figure 1.2: Device parametric summary of custom low power technology used for LPPTM simulations. Model centering based on published data from [87] [55] [54] [88]

.

mechanism and a major concern for future SRAM designs.

This local variation is best characterized by measurement of the mismatch between two identically drawn transistors in close proximity to one another. The variation in mismatch is then defined by:

$$\sigma Vt_{mm} = A_V t \cdot \left(\frac{1}{\sqrt{WL}} \right) \tag{1.1}$$

where the quantity has units of and W and L refer to the device width and length respectively. What has now become commonly referred to as the Pelgrom plot, where the delta Vt of two identically drawn adjacent devices provides an essential relationship between two essential design parameters (W and L) and the expected random variation expectation. Based on published hardware measurements [60] [94] for competitive industry technologies, the A_{Vt} value used was for the LP PTMs in this work was 3mV-μm. The channel

length variation (both global and local), and the variation in Vt associated with implant dose variations were also included. The combined effects of scaled gate oxide thickness (Tox) of approximately 10% per generation over the range of technologies included and corresponding increase in the effective channel doping (N_A) of approximately 20% per generation tend to hold the values roughly constant with each generation, which may be explained by the commonly used empirical equation [6] [5]:

$$\sigma Vt = 3.19 \cdot 10^{-}8 \cdot T_{ox} \cdot \left(\frac{N_A^{0.4}}{\sqrt{WL}} \right) [V] \tag{1.2}$$

Asenov's empirical equation, derived through atomistic simulation results, affirms the Pelgrom relationship to $1/\sqrt{WL}$ and includes the important role of the gate capacitance and channel doping. A first principles treatment was developed as early as 1975 by R.W. Keyes, relating the predicted variation to the channel area, random channel dopant fluctuations and gate capacitance [41]. Using percolation theory and simple channel doping profiles the following relationship was derived [41]:

$$\sigma Vt_{mm} = \frac{q}{2\frac{\epsilon_{ox}}{T_{eq}}} \sqrt{\pi N_A} \cdot (WL)^{-3/8} \cdot \left(\sqrt{\frac{4\epsilon_{Si}(\frac{k_B T}{q} \cdot ln(N_A/ni))}{q \cdot N_A}} \right)^{1/4} \tag{1.3}$$

where Teq is the gate oxide equivalent thickness, k_B is Boltzmann's constant, T is temperature in Kelvin, q is the fundamental charge of an electron in eV, ϵ_{ox} and ϵ_{Si} refer to the dielectric constants for the gate oxide and silicon respectively, and N_A is the doping concentration in at/cm^3. Although today's device designs are much more complex, commonly involving extension and halo implants, the relationships to gate capacitance and doping concentrations are consistent. For most empirical data in recent evaluations the inverse square root relationship to channel area rather than the inverse (3/8) power relationship as

derived by Keyes is used.

For clarity, σVt_{mm} refers to the mismatch between two identically defined devices in close proximity while the σVt for an individual device is therefore smaller by a factor of $\frac{1}{\sqrt{2}}$. The implications of the scaled devices employed in the SRAM cell is shown in Fig. 1.3. Both components assumed a 3σ value equal to 10% of the target (Lphysical) for the technology. A 30mV (3σ) global variation in Vt_0 for NMOS and PMOS due to implant dose and energy variability was assumed.

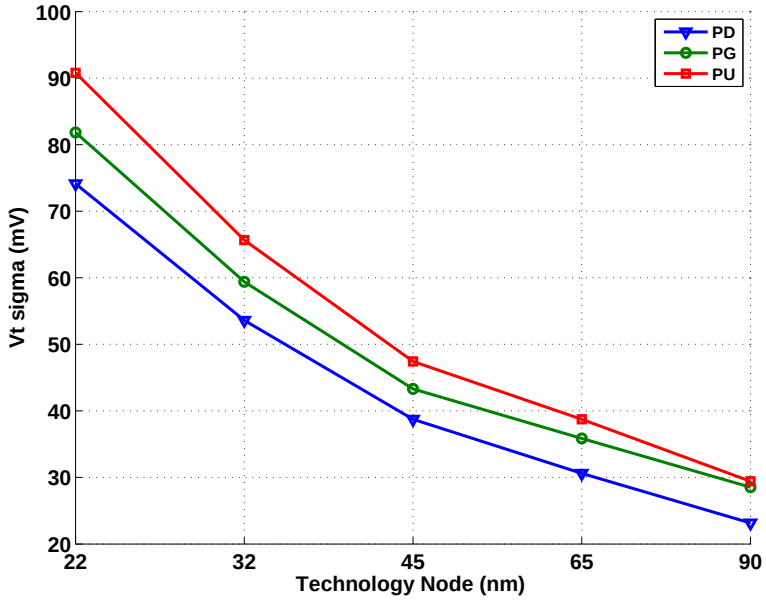


Figure 1.3: Impact of scaling trends on pull down (PD), pass gate (PG) and pull up (PU) SRAM device Vt sigma based on the RDF component.

The infrastructure resulting from this analysis coupled with the known scaling relationship for the SRAM devices from Fig. 1.1 provides a means of assessing the local variation in threshold voltage across the technology nodes. This relationship is shown in Fig. 1.3, where the local variation in SRAM device threshold voltage has increased for the 32nm node by roughly a factor of 2 over the variation addressed at the 90nm node. The adoption

of high-κ with metal gate beyond 45nm may provide some relief, consistent with (1.2), but the increasing trend will again increase as the channel area is reduced through scaling.

The A_{Vt} for emerging FDSOI technologies will be of significant interest with potential improvements in the channel dopant variation [73] [48]. To enable further exploration of this potential technology solution, the extraction of the A_{Vt} for a 150nm FDSOI technology was accomplished by the design and implementation of characterization circuits and test methodology.

In addition to the well know sources of random variation, a deeper exploration and focus was placed on looking at the potential sources of systematic variation that can arise from the combination of bit cell topology, use of pushed design rules and industry scaling practices. The sources of non-random mismatch are investigated in the context of the 6T SRAM cell layout. These systematic offsets play a role in the yield expectations of the large arrays due to the impact on the noise margin distributions. Both doping and geometrical systematic variation considerations are examined in the dense SRAM cell environment.

An analysis of the implications of the bit cell topology on non-random, within-cell variation and the evolution of lithography practices with continued scaling, a new bit cell topology is proposed. An examination of the growing lithography constraints and known bit cell layout topologies, the new topology may provide a path to enable further scaling of the 6T SRAM.

1.1.2 Reduced Functional Noise Margin

As voltage and area are reduced by continued scaling, the functional margins for all three required operations are becoming less robust. The consequences of scaling on SRAM

design and noise margins have been the subject of many investigations [14] [13] [31] [86] [93] [2]. We will refer to these margins as; write margin (WM), read static noise margin (RSNM), and hold static noise margin (HSNM). All three essential functions required of the SRAM; 1)write, 2)read, and 3) retain-state, all become more difficult at lower voltages. Additionally, as the channel area is reduced without proportionally scaling the Teq, device variation will be increased.

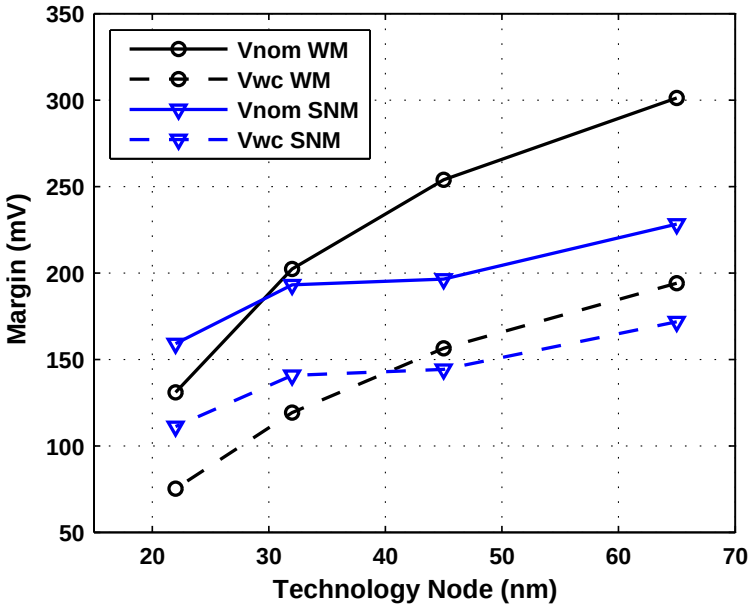


Figure 1.4: Simulated LP SRAM functional margins (RSNM and WM) are decreasing with continued technology scaling. 'Vnom WM' and 'Vnom SNM' refer to the nominal simulated margin value and 'Vwc WM' and 'Vwc SNM' refer to the worst case margins out of 1000 Monte Carlo simulations.

Fig. 1.4 depicts the simulated functional noise margins, which are trending lower with each successive technology generation. The read static noise margin (RSNM) or (SNM) is a measure of the stability of the cell during access [74]. The RSNM metric quantifies the resistance of the cell to upset during a read operation and the write margin (WM) metric quantifies the ability to write data to the cell. The SNM metric is a very critical metric

because all bits along the asserted word line will be subject to a SNM upset during a read operation as well as the bits along the unselected columns during a write operation. This is commonly referred to as the half-select issue.

To address the trend in reduced margins, a range of circuit assist methods have been proposed. Two chapters in the thesis explore bias based circuit assist methods for nanoscale SRAM. Bias based circuit assist techniques will be defined for the purposes of this work as an intentional modulation of an accessible terminal(WL,BL,VDD,VSS,Body) voltage, charge or timing with the goal of improving the read or write margin. An objective method for assessing the effectiveness of the various assist options will be developed in this work. A margin/delay analysis is developed to further improve and provide clarity for future investigations in the assist space.

Although assist methods do offer a path to extend the 6T operation window and provide yield improvements by effectively lowering the array Vmin, limitations exist. The limitations of the bias based assist methods for read access provided a unique and clear result which is provided by the assist limit contour (ALC). This ALC contour allows the circuit designer to quickly establish the limits of the bias based circuit assist methods for a given process technology.

1.1.3 Increased Standby Leakage

While the functional margins are decreasing with continued scaling, the standby power for the array, as measured by the bit cell parasitic leakage, is increasing. There are three primary mechanisms involved in this trend. First the gate tunneling current is increasing with Tox reduction [52] [47]. The tunneling mechanisms are voltage accelerated and ex-

hibit little temperature dependence. The second is gate induced drain leakage (GIDL) [96]. The use of halo or pocket implants to improve the short channel effects (SCE) by reducing drain induced barrier lowering (DIBL) has tended to increase GIDL in the devices. For low power technologies, GIDL may be a significant component of the off state parasitic leakage. The third major leakage contributor is sub-threshold leakage [81]. This mechanism is governed by the sub-threshold slope and the threshold voltage of the device by the following relationship:

$$I_{sub} = I_0 \cdot \frac{W}{L} \cdot 10^{(\frac{Vgs-Vt}{S})} \tag{1.4}$$

where I_0 is a technology dependent constant with units of current, Vgs is the gate to source voltage (0V in off state), Vt is the threshold voltage and S is the sub-threshold slope [81]. S has units of mV/decade and is expressed in the relationship (1.5).

$$S = ln(10) \cdot \frac{kT}{q} \cdot \left(1 + \frac{\epsilon_{Si}}{\epsilon_{ox}} \frac{Tox}{X_d}\right) \cdot \left(1 + \frac{11 \cdot Tox}{X_d} exp\left(\frac{-\pi \cdot L_{eff}}{2 \cdot X_d + 3 \cdot Tox}\right)\right) \tag{1.5}$$

In expression (1.5), Tox is the gate oxide thickness, k is Boltzmann's constant, T is temperature in Kelvin, q is the fundamental charge of an electron in eV, X_d is the depletion thickness, L_{eff} is the effective channel length. The threshold voltage also tends to decrease with the technology Vdd in order to achieve sufficient overdrive to preserve performance. Because GIDL and gate leakage are tunneling mechanisms and exhibit little temperature dependence, the sub-threshold leakage becomes the dominant leakage source at elevated temperatures. The technology choice is often a critical factor in the array standby power. Technology solutions optimized for low SRAM standby power have achieved leakage values averaging $< 50fA/cell$ at 25C [58].

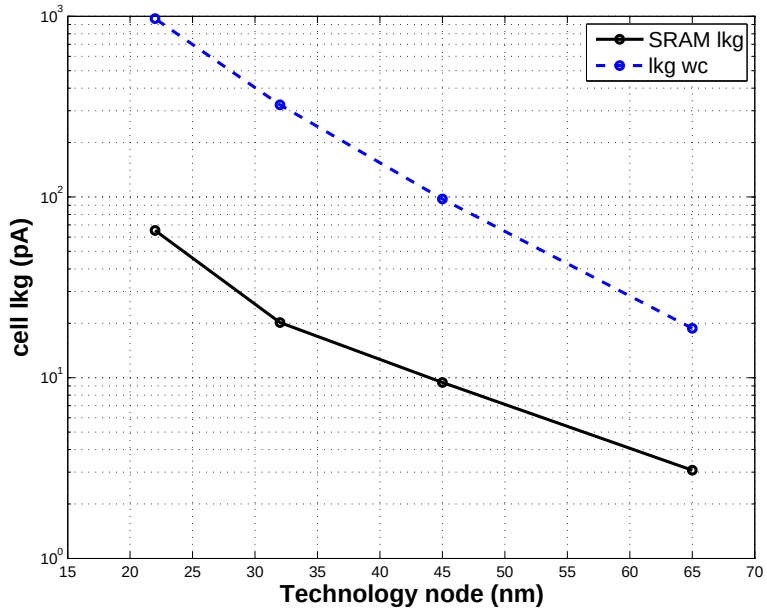


Figure 1.5: Simulated nominal and worst case (out of 1000 Monte Carlo cases) parasitic leakage per 6T SRAM cell is based on the predictive LP technologies.

As can be observed from (1.4) and (1.5), the introduction of high-κ dielectric materials for the gate dielectric can improve the sub-threshold slope by allowing reduced Tox values and therefore the sub-threshold leakage. The net contribution of the high-κ material therefore is significant in providing a path to improve variability, gate leakage, and, to some extent, sub-threshold leakage.

1.1.4 SER Susceptibility

Another important area of concern for nanoscale SRAM is increased susceptibility to radiation induced soft errors. Although this topic is not specifically developed in this dissertation, it is a issue that should be highlighted when discussing challenges faced as we continue to scale the 6T SRAM. Soft errors in the form of both single event upsets (SEU) and SRAM array multibit fails [7] [21] [75] represent a reliability concern for the memory

designer.

The two primary sources of soft error inducing radiation are from either terrestrial radiation or from radioactive isotopes within materials used in the integrated circuit fabrication process. High energy cosmic radiation interacting with the earth's atmosphere results in a flux of neutron particles with a large range of energies extending to several 100MeV [98]. At sea level the resulting high energy neutrons manifest a relatively isotropic flux of 10-20 neutrons/cm^2-hr and can interact with the silicon lattice through elastic and inelastic recoil or by spallation where the silicon atom is shattered into heavy and one or several lighter particles. This process produces a charge cloud of electron-hole pairs that, when in close proximity to one or more sensitive neighboring circuit nodes, may result in a single or multi-bit error.

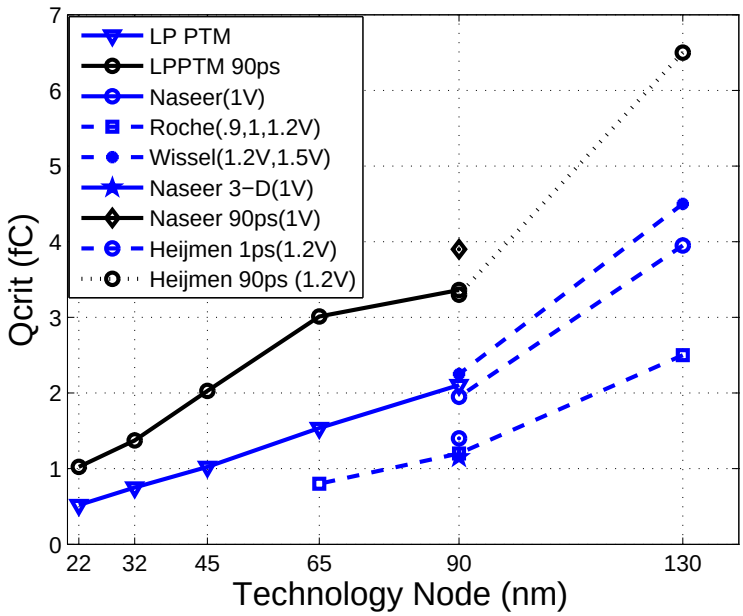


Figure 1.6: Trend in 6T SRAM Qcrit values with continued scaling.

The second form of radiation which predominately originates from impurities within

the materials used in modern interconnect technology is the alpha particle. The alpha particle can be characterized as a doubly ionized (He) atom consisting of 2 protons and 2 neutrons. The alpha particle originating from impurities found in the interconnect or packaging materials used in integrated circuit manufacturing has an initial energy extending up to 8 MeV depending on the specific impurity isotope present. Current purity levels in VLSI processing are sufficient to insure that the alpha flux is not greater than 0.001 /cm^2-hr. Because the alpha particle is ionized, it interacts with the silicon lattice to produce a column of electron-hole pairs along the path of the particle, which can cause an upset if the charge collected at the circuit node exceeds the critical charge (Q_{crit}) for that circuit or memory bit.

The soft error rate (SER) is expressed as:

$$SER \simeq F \cdot A_{diff} \cdot exp\left(\frac{-Q_{crit}}{Q_S}\right) \tag{1.6}$$

where F is the particle flux, A_{diff} is the critical or sensitive charge collection area, Q_{crit} is the critical amount of charge required to flip the bit [28] and Q_S is the charge collection efficiency. Cell design topologies that minimize A_{diff} and increase Q_{crit} are therefore preferred. The charge collection efficiency Q_S is modulated by factors such as voltage, charge sharing, NWELL and PWELL depth, use of retrograde well doping profiles and use of triple well. The amount of charge collected at a given node is typically much less than the total charge generated. Values for Q_S are obtained following the trends provided by Hazucha and Svensson [27]. Fig. 1.6 provides a summary of published Q_{crit} values as well as simulated Q_{crit} for the scaled bulk technologies defined in this work down to 22nm.

SOI technologies have been shown to offer improved resistance to soft error upset, and direct comparisons between SOI and bulk technologies show $\sim 5\times$ improvement for the

SOI [15] over bulk. Although the diffusion capacitance is much lower for SOI technologies (increasing the sensitivity), the charge collection efficiency for SOI is much smaller compared to bulk technologies.

1.1.5 NBTI and PBTI Sensitivity

An additional challenge that confronts the nanoscale SRAM design is the shift in threshold voltage during product lifetime. The most significant mechanism for this has been negative bias temperature instability (NBTI), which results in a degradation of the PMOS device associated with a shift in the threshold voltage [40] [45]. This induces a corresponding shift in the functional margins of the SRAM cell discussed earlier. The SNM will be decreased by an amount typically on the same order of the mean Vt shift of the PMOS device while the write margin will be improved by the weakened PMOS.

Although PBTI was not observed to play a significant role for technologies with conventional $Si0_2$ and nitrided oxide gate dielectrics, with the introduction on high-κ gate dielectric materials such those involving Hf oxides and oxy-nitrides, the PBTI mechanism is a renewed concern [78]. The use of NMOS devices for the access transistors, which is common in today's 6T and alternative bit cell options, this can result in degraded performance and yield impacts.

1.2 Summary

A number of obstacles exist to the continued use and scaling of SRAM designs beyond 32/28nm. These include increased variation, reduced noise margins, increased standby

leakage, and reliability detractors such as NBTI and radiation induced soft errors. Despite these detractors, new advances in technology and circuit design offer promising options that provide a path forward. This work is directed toward addressing reduced SRAM functional margins which accompany continued technology scaling. A question consistent with this theme may be expressed simply as, "What is the future of 6T SRAM beyond the 32/28nm node?" While there are many aspects to this question, this work investigates and addresses 1) systematic variation sources in SRAM devices, 2) an optimized method for selecting a circuit assist scheme, 3) the limits of bias based assist methods, and 4) a new bit cell design for future nanoscale SRAM.

1.3 Major Contributions

1. **Highlight specific sources of non-random mismatch in the context of the aggressive bit cell design environment**

   A new examination of SRAM device variation sources for the nanoscale era, a highly important aspect of advanced large scale CMOS memory design, is presented. Specifically, a description of how dopant fluctuations in nanoscale SRAM devices may be attributed to both random and non-random components. Three factors which play a role in the susceptibility to sources of non-random dopant variation are; 1) SRAM cell layout topology, 2) process scaling practices, and 3) pushed design rules used in dense SRAM bit cell designs. Both doping and geometric sources of variation are addressed.

   Four specific sources of dopant fluctuation which can contribute to non-random

threshold mismatch in the SRAM device environment are; (1) implanted ion straggle in SiO_2, (2) polysilicon inter-diffusion driven counter-doping, (3) lateral ion straggle from the photo-resist and (4) photo-resist implant shadowing. A manuscript titled "Non-random device mismatch considerations in nanoscale SRAM" has been submitted for publication to IEEE Transactions on VLSI Systems. This work is believed to be the first to highlight and address all four mechanisms of systematic dopant driven mismatch in context of the aggressive bit cell design environment.

2. **Propose a new bit cell topology for the sub-22nm era**

   As scaling continues, the lithography challenges grow and can assert changes in the layout topology of the bit cell. A new bit cell topology is proposed that offers 1) reduced metal 1 complexity, 2) eliminates jogs in the active silicon for reduced geometric variation, and 3) offers shorter M2 bit lines over the dominant industry bit cell used today. A provisional patent has been submitted on this new bit cell design topology [59]. A manuscript titled "New category of ultra-thin notchless 6T SRAM cell layout topologies for sub-22nm" has been submitted for publication to the proceedings from 12th International Symposium on Quality Electronic Design.

3. **Developed the margin/delay analysis metric for circuit assist analysis**

   The primary focus of circuit assist methods has been improved read or write margin with less attention given to the the implications for performance. In this work, margin sensitivity and margin/delay analysis tools are introduced for assessing the functional effectiveness of the bias based assist methods. A margin/delay analysis of bias based circuit assist methods is presented, highlighting the assist impact on the functional

metrics, margin and performance.

A new method for concurrently optimizing the impact of circuit assist methods and biases is presented and referred to as the margin/delay method. The concept of margin sensitivity is developed and discussed as a necessary component of the margin/delay concept. The analysis spans four generations of low power technologies to show the trends and long term effectiveness of the circuit assist techniques in future low power bulk technologies. A publication titled " Impact of circuit assist methods on margin and performance in 6T SRAM" was published in the Journal of Solid State Electronics [57].

4. **Address the limitations of bias based assist methods and highlight the value of the assist limit contour (ALC)**

   Although circuit assist schemes provide improved yield margin for scaled SRAM, factors such as reliability, leakage and data retention establish the boundary conditions for the maximum voltage bias permitted for a given circuit assist approach. These constraints set an upper limit on the potential yield improvement that can be obtained for a given assist method and limit the minimum operation voltage (Vmin). By application of this set of constraints, it is shown that the read assist limit contour (ALC) in the margin/delay space can provide insight into the ultimate limits for the nanoscale CMOS 6T SRAM. A paper titled "Limits of bias based assist methods in nanoscale 6T SRAM" was published in the proceedings from 11th International Symposium on Quality Electronic Design [56].

1.4 Organization

This thesis is constructed in the following manner: Following the background and introduction provided in this chapter, chapter 2 describes an investigation of the sources of variation (both random and systematic) for SRAM cell devices. Although much research and discussion has been given to the issues of random variation sources, such as RDF, this chapter describes four mechanisms that can be found to induce non-random mismatch in the SRAM bit cell devices.

In chapter 3 the challenges associated with future SRAM bit cell design are discussed, and the geometric variation sources which can contribute to within cell mismatch in the highly scaled array environment are examined. A new bit cell layout topology is proposed, and its attributes are examined. The advantages and disadvantages of this new cell topology for future 6T dense SRAM are identified and compared against the industry standard bit cell.

Chapter 4 presents an in depth analysis of the bias based assist techniques available for 6T SRAM. A method for categorizing the 6T SRAM assist options is presented. A margin/delay analysis technique is developed to allow the concurrent evaluation of both performance and margin for the circuit assist methods in 6T SRAM. The assist methods are explored across for technology nodes to better understand the impact of scaling on the assist benefits for future generations.

Chapter 5 builds on the foundation work presented in chapter 4 and explores the limits of bias based assist methods for nanoscale SRAM. Because factors such as reliability, leakage, and data retention establish the boundary conditions for the maximum voltage bias permitted for a given circuit assist approach. These constraints set an upper limit on the

potential yield improvement that can be obtained for a given assist method and limit the minimum operation voltage (Vmin). By application of this set of constraints, it is shown that the read assist limit contour (ALC) in the margin/delay space can provide insight into the ultimate limits for the nanoscale CMOS 6T SRAM.

Chapter 6 summarizes the main contributions presented in this thesis and discusses potential direction for future work building on this work.

Chapter 2

Variation: Sources of random and non-random device mismatch in nanoscale SRAM

2.1 Introduction

The SRAM cell area has become a benchmark of technology competitiveness in today's VLSI industry. The design trade-offs to achieve the aggressive SRAM bit cells are becoming more challenging with each successive technology generation. To achieve the density, performance, and functional requirements, the competitive bit cell requires design rules which are much more aggressive than those used in base logic designs. For this reason, the bit cell has become an integral part of the technology offering for technology suppliers. Because SRAM is largely compatible with CMOS logic processing and failures can be readily identified through bit fail mapping, it is commonly used by industry as a technology quali-

fication vehicle. Although many design rules are limited directly by lithography, there are several mechanisms which must be addressed for the commercially successful nanoscale cell design.

2.2 Background and Motivation

The complex set of decisions that must be addressed in defining the competitive SRAM bit cell design require a combined understanding of device physics, process integration capabilities, as well as an understanding of the circuit and memory architecture design. Additionally, commercial success will require an understanding of the competitive market as well in order to optimally balance the density, performance, functional margins and power requirements. In this chapter, the interaction of process integration technology in SRAM cell design is explored using technology computer aided design (TCAD) process simulation tools. The Synopsys simulation tools and simulation environment will be used to specifically examine the scaling limitations and challenges for the SRAM cell design.

A significant source of variation in nanoscale CMOS technologies is associated with random dopant fluctuations (RDF), which follows a $1/\sqrt{WL}$ relationship. Although high-κ/metal gate technologies have provided some relief, aggressive design rule and device scaling has led to an increase in device variation in both SRAM and logic devices. Because it is common for the SRAM devices to be near or below minimum logic design rules, the RDF mismatch phenomenon is exacerbated. Additionally, pushed design spacing rules used in the dense SRAM cell can lead to added sources of variation that is not observed in circuits designed with the standard logic design rules.

Although the SRAM devices and logic devices are built concurrently using the same

processing steps, often sufficient differences exist so that separate BSIM device models are required for the SRAM devices. This may be attributed to several factors. First, there may be intended deltas due to the use of additional Vt tailor steps to fine tune the SRAM threshold voltage for optimal functional (yield) margin, performance, or leakage optimization reasons. The second reason is non-intentional and is attributed to the process, structural differences, STI stress, and a range of proximity effects. For these reasons, commercial nanoscale CMOS technology suppliers provide a set of unique models for the SRAM cell devices that accompany the supplied bit cell. These additional sources of variation may also contribute to non-random or systematic mismatch within the SRAM cell.

As scaling continues beyond the 32nm node, the pushed rules used in bit cell design will warrant increased attention and more costly measures to avoid sources of systematic, non-random device mismatch. We define non-random mismatch as a mean offset in the device pair (e.g. pull down NMOS Vt left vs right) within the same or adjacent bit cell. Factors that may contribute to non-random mismatch are layout topology, process scaling practices, and use of pushed design rules in the bit cell.

In this chapter, the implications of cell layout topology, process scaling, and pushed design rules are considered. Four specific alignment sensitive mechanisms which may impact non-random device threshold mismatch are evaluated. Experimental data and process simulations are used to both highlight and quantify sources of non-random mismatch. A statistical basis is provided as a foundation for quantifying the functional margin impacts of non-random device variation on the bit yield. Based on an examination of existing 6T layout options, and consideration of non-random mismatch sources, we examine the relative merits of an alternative layout, possessing different symmetry, and area limiting design

rules from the topology used in today's dominant industry layout.

Four sources of potential non-random threshold mismatch that can arise from the use of aggressive design rules in the bit cell are; (1) implanted ion straggle in SiO_2, (2) polysilicon inter-diffusion driven counter-doping, (3) lateral ion straggle from the photo-resist and (4) photo-resist implant shadowing. Using simulation and hardware measurements, we quantify the device parametric impacts and provide a statistical treatment forming the basis for quantification of the functional margin impacts on the bit cell. We examine two lithography compliant bit cell layout topologies and quantify the impact of systematic mismatch on the margin limited yield.

2.2.1 Cell Topology

The choice of cell topology is perhaps the most critical choice and must be made early in the technology development phase. This choice will significantly impact the ultimate cell size and aspect ratio that can be obtained as well as compatibility with assist methods. It influences the bit line capacitance and the design rules that will need to be pushed and if any unique (non-logic based) features will be desired such as the shared contact used in the majority of today's 6T bit cells. Included in the topology decision are the number of transistors to be used and the alternative bit cell options. The term "alternative bit cell" is used to describe a range of bit cell options that include the total transistor options of 4 through 10 (excluding 6T).

2.2.2 Non-6T SRAM (alternative bit cell options)

With the recognition that achieving the performance, yield, and leakage targets with the 6T cell is becoming increasing more difficult to with each technology generation, alternative (non-6T) cell topologies have been proposed. The alternative cell designs tend to provide a solution that addresses one or more of the challenges highlighted in chapter 1. A few examples of the alternative cells are: 5T which offers a path to improve stability but requires a write assist [63], the 7T [4], there are several implementations of 9T [50] [51] and 10T [11] [67] [65]. Although the alternatives do tend to provide partial solutions, the 8T cell topology [16] is becoming more commonly used in commercial applications, particularly for L1 and L2 cache applications. While the 8T cell area is larger than the 6T, it does offer several advantages. With the two added transistors as a read buffer, the read disturb mechanism is avoided (with additional architecture constraints to avoid the half select concern when the write word line is asserted). Because this design is still subject to the half-select disturb during a write operation, array architecture changes are used to avoid this mechanism. Further, this design offers both read and write word lines so additional performance gains can be realized by optimizing the read and write paths independently [17].

2.2.3 6T cell topologies

The optimal 6T layout topology will be dependent on many factors. These include processing capability, performance, density, power, and functional requirements. There are, at least initially, a number of options theoretically available for placing 6 transistors to perform the desired function. A summary provided by Ishida is reproduced in Fig. 2.1 for

this discussion [37]. Following the nomenclature of Ishida, although published examples of type 2 and 3 can be found, the cell type 1a was the dominant industry topology across several nodes prior to 90nm. At technology nodes below 90nm, the type 4 cell topology became (and remains) the dominant industry cell design.

2.2.4 Lithographic considerations

As scaling continues below 90nm, the lithographic challenges in printing and controlling the dimensions within the same printed layer in orthogonal directions has become increasingly difficult [38]. This has led to restrictions in layout orientation and shape for printed layers requiring tight dimensional control. For the SRAM devices it is therefore advantageous for the active single crystal regions and gate layer to be printed orthogonally, thus allowing optimal dimensional control for these layers. Of the cell topologies or types summarized by Ishida, only type 4 and a variation on type 1-b provide this advantage. For this reason we will explore both design topologies in more detail. Fig. 2.2 compares the dominant industry layout (type 4) with an alternate style (1x), Fig. 2.2(b), that also complies with the layout restrictions in sub 90nm designs.

2.2.5 Bit cell dimensions

Because of the design symmetry for both topologies, the bit cell area can be expressed as the product of the cell boundary dimensions $Xcell$ and $Ycell$.

$$Acell = Xcell \cdot Ycell \tag{2.1}$$

A list of the limiting pushed design rules representative of those used in advanced sil-

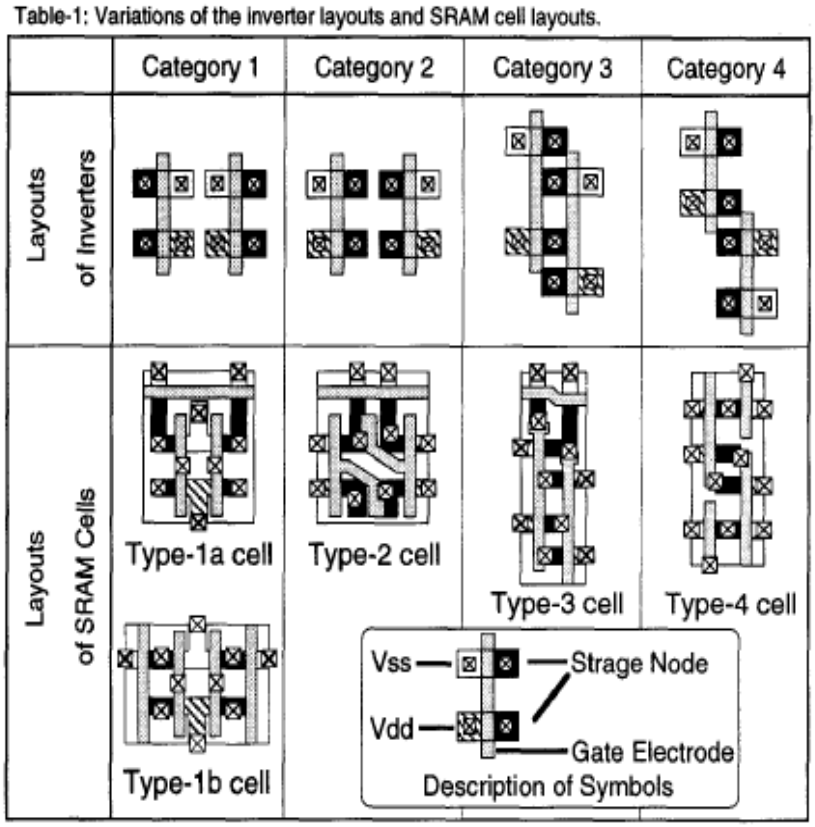
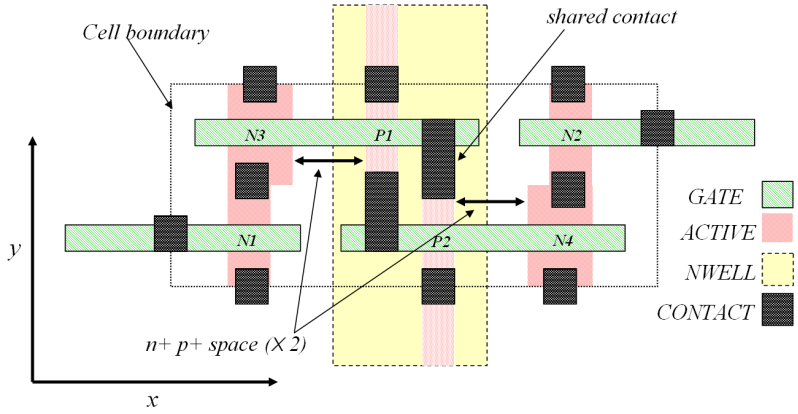
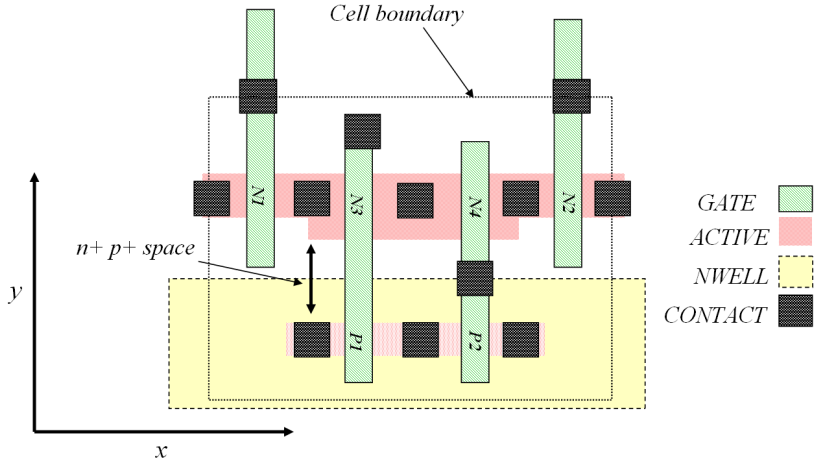


Figure 2.1: Summary of 6T cell layout topologies (©IEEE '98) [37].

icon dense SRAM cells is given in Table 2.1. Although some deviation will be expected as technologies evolve, the rules are expressed as function of the technology node (λ) to capture the effect of scaling. Although these pushed rules are consistent with those used in industry, some differences will exist between technology suppliers to allow optimization of yield and parametric values as desired. We use (Wpd/Lpd),(Wpg/Lpg),(Wpu/Lpu) to refer to the width and length of the pull down NMOS, pass gate NMOS and pull up PMOS devices respectively. The dimension (X_4) for design topology 4, illustrated in Fig. 2.2(a), becomes:



(a) Dominant industry bit cell design (topology 4).



(b) Alternate bit cell design (topology 1x).

Figure 2.2: Example layouts of 6T SRAM bit cell topologies 4 (a) and 1x (b). Alignment of NWELL layer and subsequent block level layers will be asymmetrical with respect to devices N1, N3 and P1 compared with devices N2, N4 and P2 for topology 4.

Table 2.1: SRAM bit cell design rule scaling assumptions

Design rule	**symbol**	**Dimension (λ)**
Gate to contact space	(GC)	0.7
Gate past active	(GPA)	1
Gate tip to tip	(TT)	1
Gate contact to active	(GCA)	1
Contact size	(CW)	1.4
Contact space	(CS)	1.4
p+ to p+ space	(AA)	1.7
n+ to p+ space	(NP)	1.8
M1 pitch	(M1P)	2.8

$$X_4 = 2 \cdot (\frac{1}{2}(TT) + (GPA) + max(Wpd, Wpg) + (NP) + Wpu + \frac{1}{2}(AA)) \tag{2.2}$$

and the dimension (Y_4) is:

$$Y_4 = 2(CW) + 4(GC) + max(Lpd, Lpu) + Lpg \tag{2.3}$$

Following the substitutions provided in Table 2.1, the bit cell area for topology 4 is expressed as a function of device dimensions and technology node dimension:

$$A_4 = (8.3\lambda + 2 \cdot max(Wpd, Wpg) + 2Wpu) \cdot (5.6\lambda + max(Lpd, Lpu) + Lpg) \tag{2.4}$$

We identify an alternative (1x) topology, which also conforms to the lithographic constraints previously discussed is shown in Fig. 2.2(b). While the type 1b proposed by Ishida

Table 2.2: SRAM bit cell device dimension scaling assumptions

Cell device	**symbol**	**Dimension (λ)**
Pull down NMOS width	Wpd	2.5
Pull down NMOS Length	Lpd	0.9
Pull up PMOS width	Wpu	1.4
Pull up PMOS Length	Lpu	0.9
Pass gate NMOS width	Wpg	1.7
Pass gate NMOS	Lpg	1.1

would not be consistent with the lithography constraints of today for the active silicon, the simple modification we refer to as 1x would be more preferred. Following the limiting design rule analysis as before:

$$X_{1x} = 2 \cdot (2(CW) + 4(GC) + Lpg + Lpd) \tag{2.5}$$

$$\begin{aligned} Y_{1x} = GPA + \tfrac{1}{2}(TT) + max(3.75(M1P), \\ (GCA) + \tfrac{3}{4}(M1P) + Wpd + (NP) + Wpu) \end{aligned} \tag{2.6}$$

and by substitution, Table 2.1, simplifies to:

$$\begin{aligned} A_{1x} = (11.2\lambda + 2Lpg + 2Lpd) \\ \cdot(1.5\lambda + max(10.5\lambda, Wpd + (NP) + Wpu)) \end{aligned} \tag{2.7}$$

Scaled device dimensions are approximated in Table 2.2 to enable a numerical area estimate. Although the scaled device dimensions will vary, depending on the specific fabricator device characteristics and bit cell performance and leakage targets, the measured differences in device dimensions make up a relatively small component of the overall bit

cell area. We define the device packing factor (DPF) as the total channel area of the six transistors divided by the cell area to give a measure of the efficiency of the bit cell design for a given set of device dimensions. A larger DPF implies a more area efficient cell design. For today's competitive 6T dense SRAM the DPF is on the order of 9%. The DPF for the type 1x (assuming common device dimensions) is 6%.

A graphical summary of published bit cell areas from 90nm to 22nm is provided in Fig. 2.3. The calculated area for type 1x (equation (2.7)) is approximately 50% larger than the type 4 (equation (2.4)). This is due, at least in part, to two reasons, 1) the design rules used in this analysis (consistent with those in use today) are optimized for type 4, 2) the shared contact feature, allowing a very effecient cross couple interconnection in type 4. The Y_{1x} dimension is limited by the metal 1 and contact rules, resulting in a larger $Ycell$ dimension than otherwise required given the device widths and n+ to p+ spacing given in Tables 2.1 and 2.2. Also, these rules have evolved and been optimized for the topology 4 design and will therefore tend to skew a direct comparison of area in favor of this common industry topology. If the design rules were more tailored for the 1x topology, this gap in area may be reduced.

The predicted cell area based on the pushed design rule and device scaling factors given in Table 2.1 and 2.2 show a good fit down to 32nm for the type 4 cell topology. Although only one published value is found for 22nm [26], the area of this type 4 cell design is becoming closer to the area expectation for the type 1x design. As we discuss later, the 1x topology offers some advantages over the type 4 for process complexity and susceptibility to the sources of non-random mismatch associated with the more aggressive n+ to p+ space. Based on the scaled rules given in Table 2.1 and 2.2, the type 4 topology offers improved

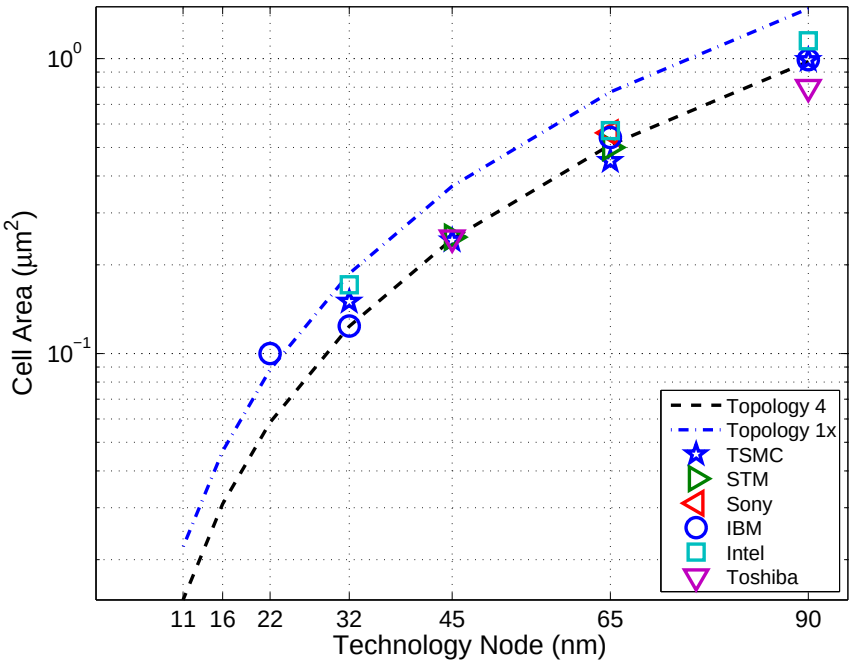


Figure 2.3: Dashed lines show SRAM bit cell areas by technology node for topology 4 and 1x based on scaled design rules and device dimensions given in Tables 2.1 and 2.2. Published 6T cell areas by technology node are beginning to deviate from the values predicted by (2.4) at 32 and 22nm.

density and, because the bit lines extend in the y-direction for both designs, a shorter (and lower capacitance) bit line (BL).

2.2.6 Process features

The shared contact feature used in topology 4 facilitates a lower DPF compared to 1x, and significantly improves the area efficiency of the cross coupled connection. While providing an advantage in area and DPF, the shared contact feature does add a degree of processing and lithography complexity above that of a logic-only process. Because this feature is typically only allowed in the well-controlled dense SRAM environment, a degree of commonality with pure logic processing is lost with topology 4, while the topology 1x

is compatible with a logic-only process and does not require this process feature.

In addition to the shared contact process feature, differences may also arise from the fact that the device pairs reside in separate active silicon islands for the topology 4 design. Although there are several potential consequences of this difference, a unique behavior in radiation induced soft error response has been observed when the separate silicon islands also share separate wells, e.g., the use of a triple well environment [23]. In contrast, the active silicon islands are shared for the device pairs for the 1x topology.

2.3 Scaling and the characterization of local random variation: device mismatch

Pelgroms method of applying a Fourier analysis to separate the global variation sources from short range (mismatch) sources such as random dopant fluctuations (RDF) can be employed to sort out random and non-random mismatch components. To perform this analysis requires a unique set of structures. The layout must be carefully controlled so any layout or local environment dependencies are minimized between adjacent pairs of identically designed NMOS or PMOS transistors. The device pairs must be drawn with a sufficient range in W and L values to enable a slope (A_{Vt}) to be extracted. These baseline structures should be logic rule based, avoid proximity to resist edges and avoid potential lithographic related shape modifications due to effects such as corner rounding, resist implant shadowing or line end foreshortening. Direct comparison of extracted A_{Vt} from the SRAM devices to that from the ideally drawn, will provide a means of assessing the degree to which the systematic variation sources discussed in section 2.4 are present. Cell topologies which reduce

Table 2.3: MITLL 150nm ULP FDSOI Technology Summary

Feature	**MITLL 150nm Technology**
Vdd	1.5V nom
Tox	4nm
Si on insulator thickness	40nm
Insulator thickness	400nm
Gate stack	20nm TiN/200nm Poly
Metal layers	3
Drawn Lmin	150nm
Silicided diffusions	13Ω/square

the systematic sensitivities while maintaining the layout density and manufacturability advantages are clearly desired.

A potential technology direction to provide improved A_{Vt} is the use of a fully depleted silicon on insulator (FDSOI) technology. To enable characterization of the benefits of this technology, a layout and circuit implementation of the mismatch structures has been completed in a 150nm FDSOI technology fabricated at MIT Lincoln Labs (MITLL). A description of the full chip and accompanying die photo are included in appendix A. A high level technology description is provided in Table 2.3.

2.3.1 Experimental method

A schematic of the circuit design, which uses a 6x64 decoder to access 8 banks of NMOS and PMOS device pairs with the following (W/L) geometry matrix: W= 5, 1, 0.8, 0.6 μm and L= 0.15, 0.2, 0.6 and 1.5 μm is shown in Fig. 2.4. The independent control

of the gates that are held off and the single gates that were swept to modulate the Vgs permitted additional control of the off state bias. A sample of the Vgs sweep results is shown in Fig. 2.5 with the off-gate held at 500mV. The off state current is comprised of the contributions of 32, 5μm wide devices with 4 different L values. In order to extract the threshold values, it was required that the off state leakage be sufficiently below the threshold current value. This was easily verified by examining the ID-VG sweeps shown in Fig. 2.5. For the 5μm wide devices the threshold currents were in a range of 0.3333e-6A for an L=1.5μm to 3.333e-6A for L=0.15μm. The layout was constructed so that the pair of devices would share a common source circuitry at M1 and above to minimize any additional sources of variation.

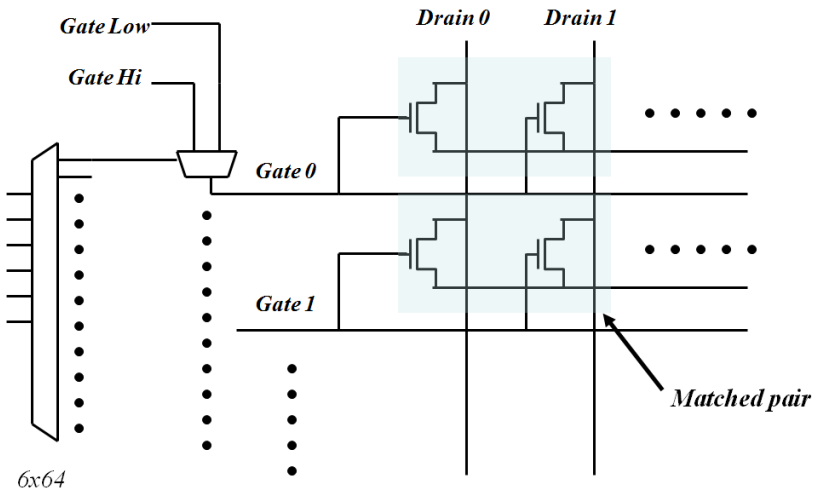
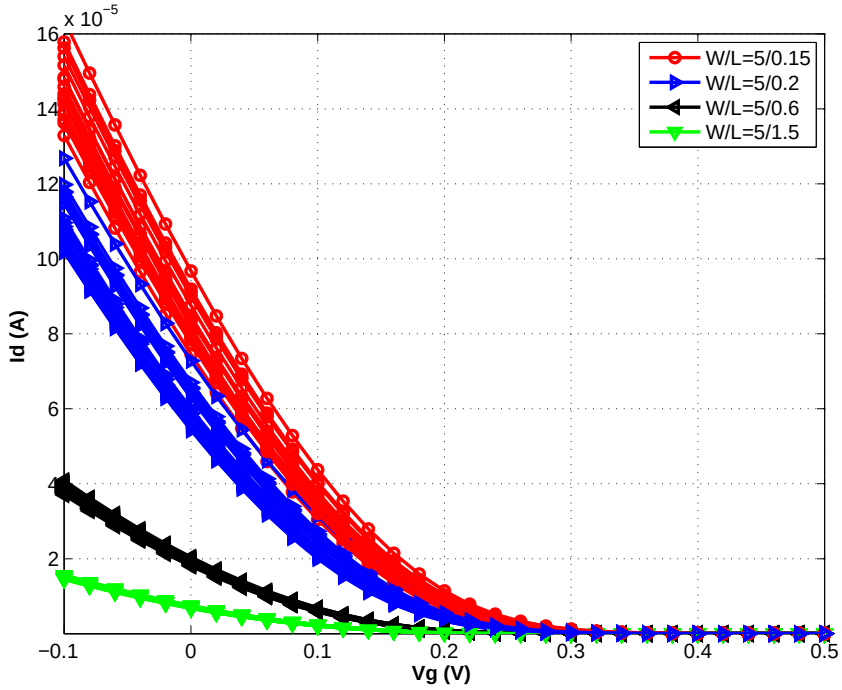
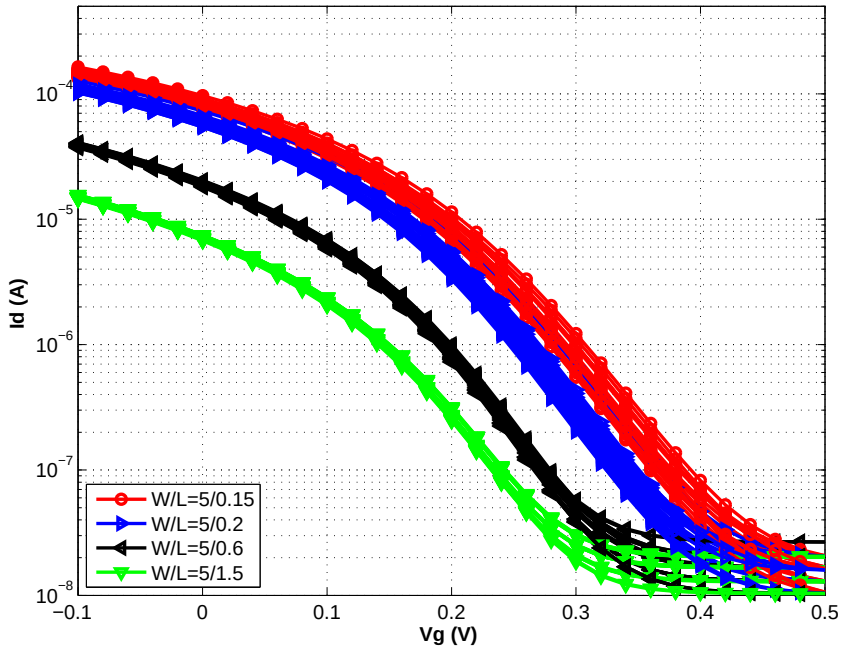


Figure 2.4: Schematic circuit diagram of device mismatch characterization circuit implementation to enable investigation of FDSOI 150nm devices.

The testing was performed using Labview 9 to control or interface with the instruments needed and to record the data taken. A Labview block diagram is included in appendix B. In

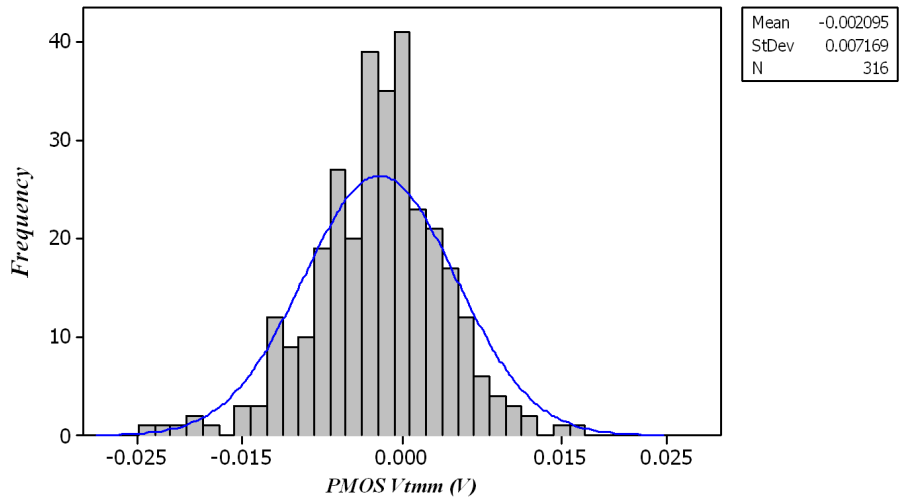


(a) Drain current vs Vgs for sample of 5μm wide devices.

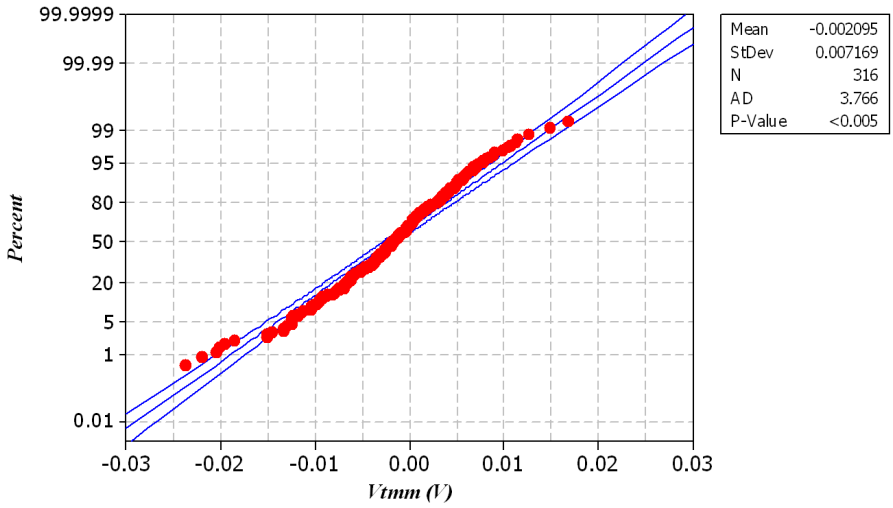


(b) Log of drain current vs Vgs for sample of 5μm wide devices.

Figure 2.5: Measured drain current versus Vgs bias for sample of 5μm wide PMOS devices.



(a) MITLL FDSOI 150nm PMOS Vt mismatch histogram.



(b) Probability plot of the PMOS Vt mismatch.

Figure 2.6: Distribution of measured Vtmm values is normally distributed and centered near zero.

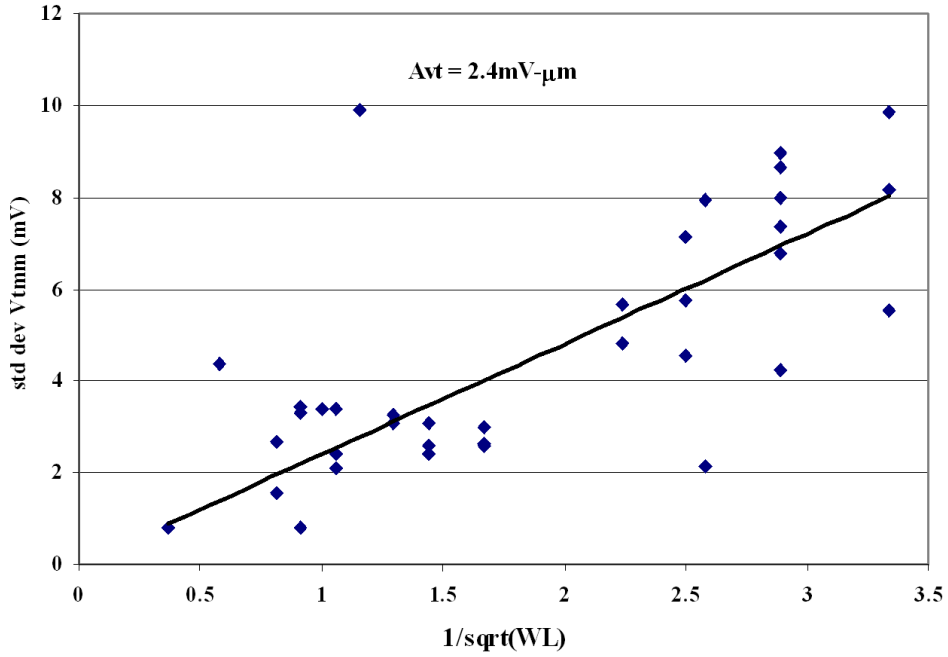


Figure 2.7: MITLL FDSOI 150nm A_{Vt} derived from the PMOS devices is 2.4mV-μm.

addition to a HP 3630A DC power supply, three instruments (Tektronix TLA7012 pattern generator, Keithley 6485 picoammeter, Keithley 2400 source meter) were used. 500mV was used for the drain to source voltage with 0V was applied to the backplate for the measurements shown. To verify the mismatch expected or mean value is centered at or very near 0V, a compiled set of measurements is shown in Fig. 2.6(a). The threshold voltage was extracted using the single point method using 100nA $\times$ (W/L).

2.3.2 A_{Vt} for FDSOI PMOS

In Fig. 2.7, the A_{Vt} for the 150nm FDSOI technology is derived from the slope of the standard deviation of the mismatch values. For this technology the PMOS A_{Vt} was deter-

mined to be 2.4 mV-μm. A value of less than 4mV-μm represents an improvement over conventional bulk technology expectations [20]. This represents a significant improvement in the random variation component of the FDSOI technologies over bulk CMOS technologies and a primary motivation for further exploration of this technology for future technology nodes.

While technology solutions are adopted and continue to be evaluated to address the random device variations that accompany scaling, SRAM devices are also subject to systematic or non-random components of variation. The next section addresses sources of non-random device mismatch in the SRAM environment.

2.4 Scaling and sources of alignment sensitive mismatch in dense SRAM

Because of its advantage in density, the type 4 topology remains the dominant cell design in the industry and has been successfully migrated across technology nodes (90nm to 32/28nm) in both bulk and SOI. However, a careful investigation of the ramifications of the continued scaling of this cell topology is warranted. In addition to added processing complexity associated with the shared contact for this cell topology, the n+ to p+ space is a cell area-limiting rule and appears twice in the cell (x) dimension, Fig. 2.2(a). As a result, the design rules associated with this space are aggressively pushed. We will now discuss how this can lead to a higher sensitivity to sources of non-random or systematic mismatch.

Reduced dimensions required for the sub-DRC SRAM bit cell as scaling continues below 32nm will place increased demands on the alignment and printed dimension tolerances

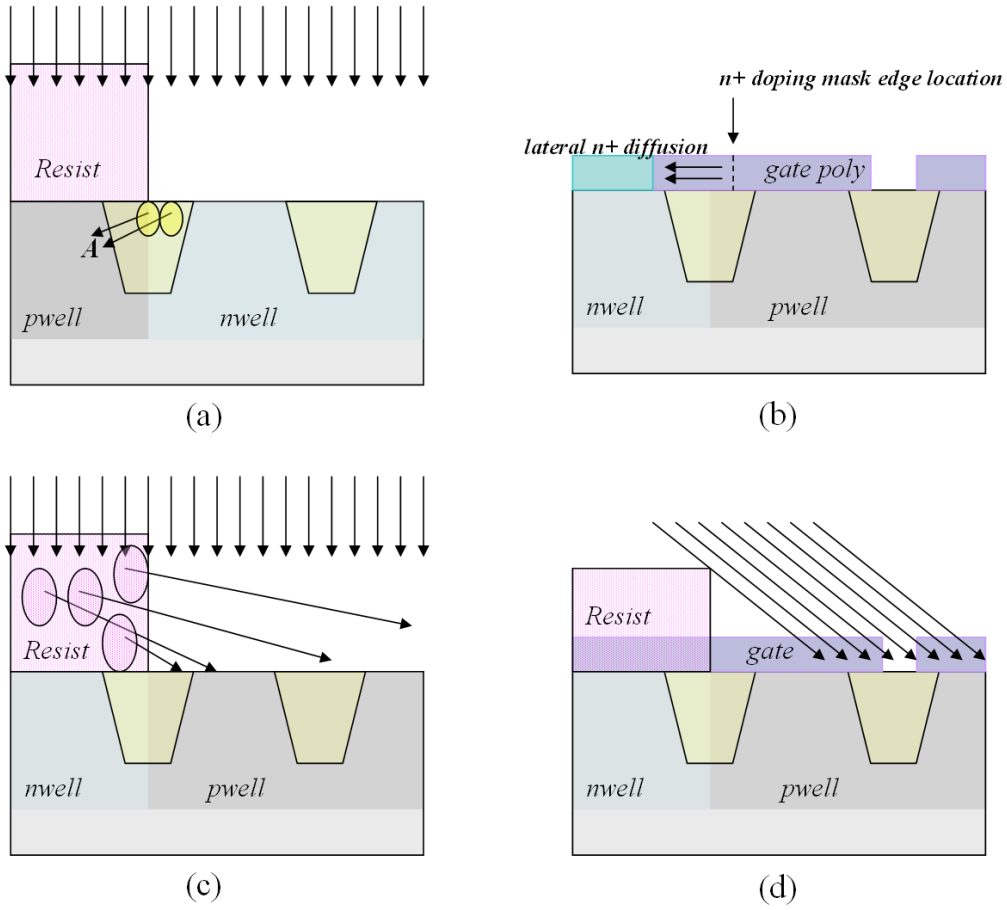
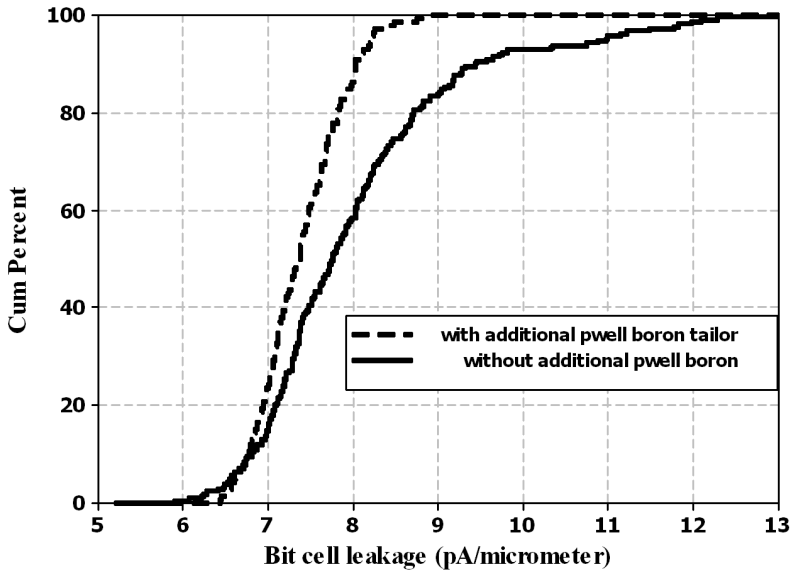


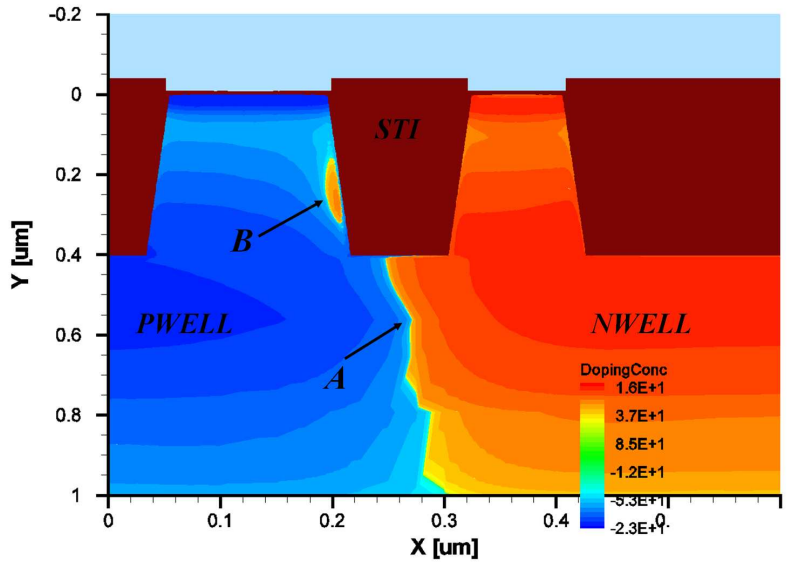
Figure 2.8: Schematic depiction of four alignment sensitive sources of potential non-random mismatch in SRAM devices. (a) Lateral straggle within SiO_2, (b) lateral counter-doping in gate polysilicon, (c) lateral straggle from resist sidewall, (4) halo shadowing.

required for large scale SRAM arrays. This is because of the yield sensitivity to the mismatch in device threshold, and any systematic mean shift or variance which is non-random ($\mu_{Vtmm} \neq 0$) will impact the soft fail limited yield.

The potential for alignment related mismatch sources is an important consideration in future bit cell design. This can arise from several factors, and the type 4 topology, while possessing a significant DPF advantage over the alternatives, is particularly vulnerable to this issue for reasons previously discussed. Fig. 2.8 illustrates four alignment driven



(a) Measured cell leakage cumulative distribution based on 24k array showing increased leakage due to lateral straggle of phosphorus in SiO_2 during NWELL implant.



(b) Simulated PWELL counter-doping due to lateral straggle of phosphorus in SiO_2 during NWELL implant with 30nm mask misalignment.

Figure 2.9: (a)Measured electrical impact on 65nm SRAM 24K array leakage due to lateral straggle of NWELL phosphorus in the STI. (b) Simulated well contours showing effects of transverse straggle in SiO_2 on the adjacent PWELL with 30nm misalignment of the NWELL resist using 45nm pushed rules. Area labeled A is normal PWELL/NWELL boundary, area B is counter-doped (n-type) region in PWELL resulting from phosphorus lateral implant straggle in STI.

sources that can introduce non-random sources of mismatch, (a) transverse or lateral straggle in SiO_2 [80], (b) polysilicon inter-diffusion driven counter-doping [53] [72], (c) lateral ion straggle from the photo-resist [30] [76] [71], and (d) photo-resist implant shadowing [32] [34]. Of these four mechanisms, (a) and (c) originate from higher energy well formation implant conditions used in bulk CMOS processes, while (b) and (d) are consistent with both bulk and SOI process technologies. In the following sections we investigate these mechanisms and their impact on the SRAM devices. Hardware data and process simulations are used to quantify the extent of mismatch from each of the mechanisms.

2.4.1 Lateral straggle in SiO_2

The potential impact of transverse straggle in the SRAM cell devices arises from the aggressive n+/p+ space used in the cell to gain density. Lateral ion scattering in the shallow trench isolation (STI) oxide from the higher energy well implants can counter dope the adjacent well edge (e.g. point A in Fig. 2.8a). The PD NMOS and PU PMOS devices are most likely to be impacted due to their proximity to the well edge. Fig. 2.9(a) shows the measured impact on the average bit cell leakage, as measured in a 24k-bit array. Although several methods of avoiding this mechanism can be taken, introducing an additional boron implant into the PWELL at the appropriate depth, $\approx$ 100nm in this case, can be used to mitigate the electrical impacts.

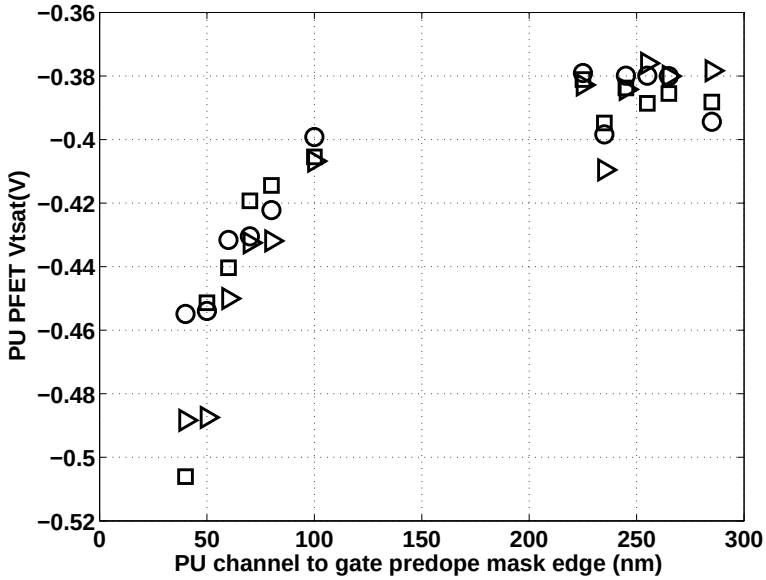
Using dimensions and implant profiles consistent with 45nm designs (n+ to p+ space of 90nm) an NWELL mask misalignment of 30nm is sufficient to create a substantial counter-doping path between the source and drain of the adjacent PD NMOS device, Fig. 2.9(b). As scaling continues beyond 45nm, the well profiles in bulk technologies will require op-

timization along with aggressive well alignment and image size tolerances to prevent this mechanism from impacting future SRAM devices.

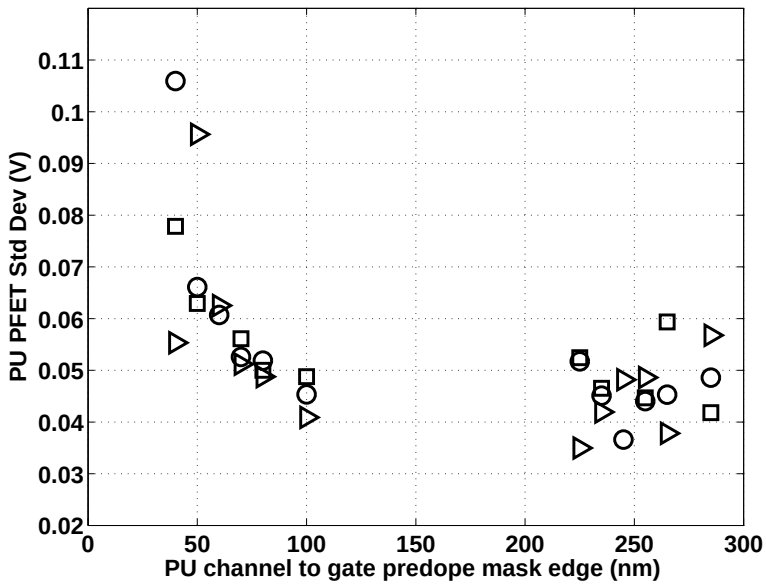
This mechanism will impact devices only on one side of the cell for the type 4 topology, thus creating a non-random device mismatch within the cell. Because the n+ to p+ space is not a limiting rule for the type 1x cell, this mechanism is much less likely to be a concern. Due to the symmetry properties of the type 1x cell, if lateral ion straggle in the STI were to penetrate into the opposite polarity well, non-random mismatch would be observed between devices in adjacent bit cells.

2.4.2 Polysilicon inter-diffusion

Although migration to metal gate began to occur at the 45nm node, some fabricators have opted to remain with polysilicon gate electrodes [89]. Polysilicon inter-diffusion is also of significant concern with scaling as n+/p+ space is aggressively pushed. The practice of using a poly pre-doping step is commonly used to insure the n+ polysilicon is degenerately doped. The alignment of this pre-doping mask as well as the n+ and p+ source drain implant masks must be carefully placed to avoid diffusion induced counter-doping of the gate above the channel region of the complementary device as shown in Fig. 2.11. Because the diffusivity is significantly higher along the grain boundaries in the polysilicon than in the single crystal, and because of the proximity of physical layout, gate counter-doping can occur. Scaling the lateral dimensions without reducing the thermal budget or alignment tolerance and/or bias will increase the sensitivity to the mechanism with scaling. Fig. 2.10(a) shows the shift in PMOS Vt_{sat} as a function of proximity to the gate predoping mask. Because the gate workfunction is also impacted, the standard deviation of the PMOS



(a) Measured PU PMOS $Vtsat$ vs gate predoping mask proximity (distance B to C in Fig. 2.11). PMOS $|Vt|$ increases as the poly predoping mask edge (C) becomes closer to the PMOS channel edge (B).



(b) Measured PU PMOS $Vtsat$ standard deviation vs gate predoping mask proximity (distance B to C in Fig. 2.11). A pronounced increase in Vt variation is observed as the gate predoping mask edge (C) becomes closer to the PMOS channel edge (B).

Figure 2.10: Effect of proximity to gate predoping mask edge on (a) PU PMOS Vtsat (b) PU PMOS Vt standard deviation. Measured data from 65nm process technology where symbols represent values measured from separate wafers.

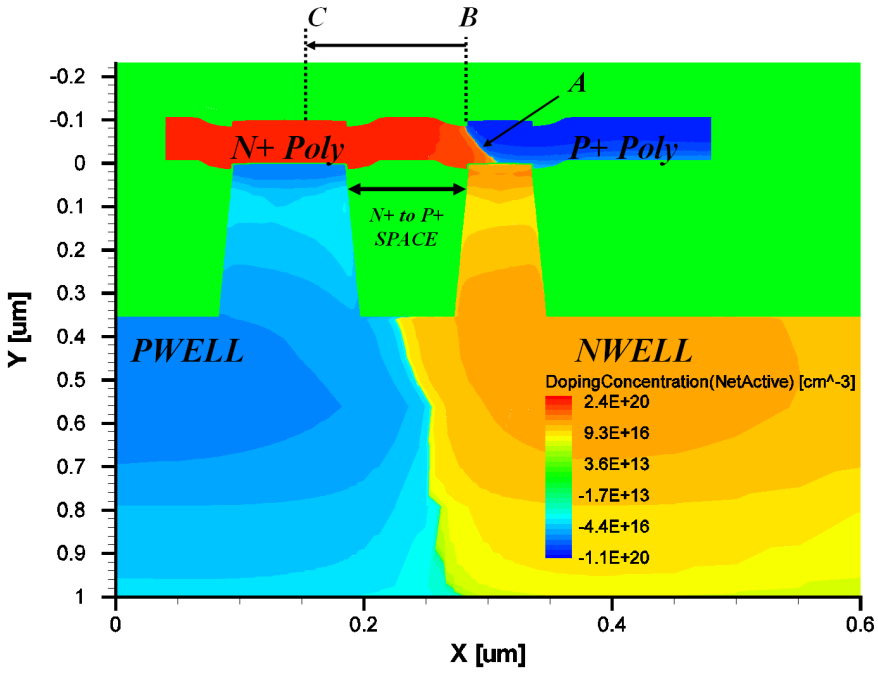


Figure 2.11: Cross section simulation illustrating the concern with poly inter-diffusion across the narrow n+/p+ space in the dense SRAM environment with type 4 cell topology. Region A shows the phosphorus encroachment over the channel region of the pull up PMOS device altering the PMOS gate work function and threshold voltage (μ, σ).

Vt_{sat} increases with phosphorus encroachment in the gate over the PMOS channel region, Fig. 2.10(b).

For the type 4 layout topology, this mechanism can result in an asymmetric mean shift in the pull up PMOS Vt as well as an increase in the variance of the Vt due to the impact on the work function component of the variance in the threshold voltage as described in (2.13). Because of the inherent sensitivity of this mechanism to the n+ to p+ space, the 1x topology would provide relief in allowing a more relaxed alignment and image tolerance specification.

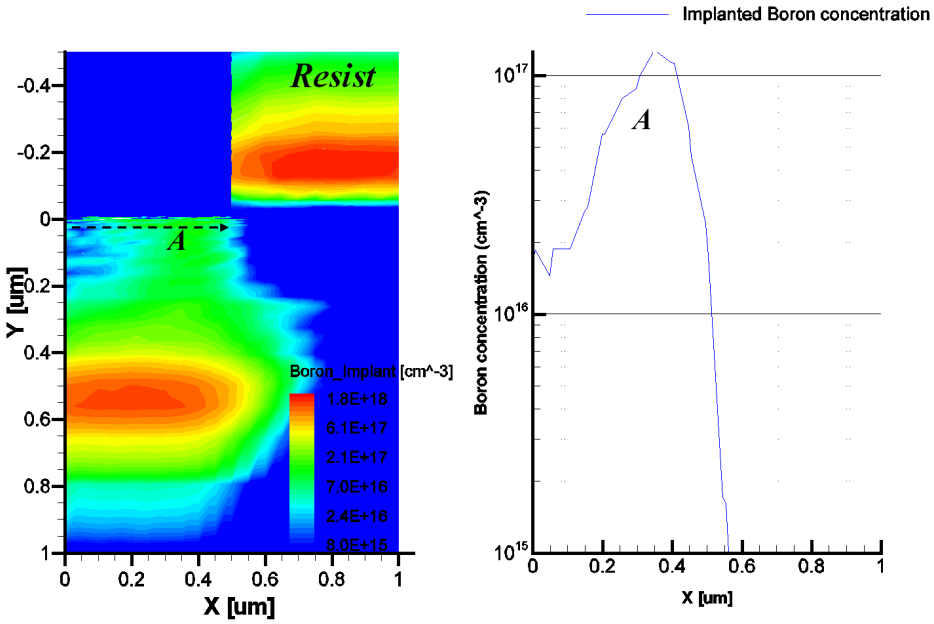


Figure 2.12: Doping contour plot following an atomistic Monte Carlo simulation of the PWELL deep implant (left). Variation in boron concentration across the silicon surface as a function of proximity to resist edge (right). Doping profile taken at a depth of approximately 50nm. The resist is located from 0.5μ to 1μ on the X axis. Boron lateral straggle emanating from the resist sidewall region during deep PWELL implant results in near-surface concentration variation across the PD NMOS channel region (A).

2.4.3 Lateral ion straggle from the photo-resist

The physical mechanism of lateral dopant straggle stemming from nuclear collisions of the high energy implant species in the photo-resist has been previously documented [30]. Depending on the implantation species and acceleration energy, this mechanism can impact devices in proximity to the well edge at distances exceeding 1μm. Because of this, this mechanism can impact both logic devices as well as the devices in the dense SRAM cell. For bulk technologies requiring higher dose and energy well implants the effect is more significant.

The amount of near surface doping is proportional to the dose of the high energy implant

used in the formation of retrograde wells. As shown in Fig. 2.12, using implanted B11 energy of 200keV with a dose of 3E13 at/cm^2, the near-surface doping is a function of the distance from the resist sidewall.

Because the surface concentration is a function of the distance from the resist sidewall, there is an alignment sensitivity for the SRAM devices. Because of the higher channel doping levels and use of thin oxide devices used in most nanoscale SRAM cells, provided deep retrograde implant doses are kept in this range, the impact of this mechanism on nanoscale CMOS SRAM is expected to be limited.

Because of the dense bit cell layout requirements, the SRAM devices in bulk technologies are subject to this mechanism regardless of the cell topology. The implications of this proximity mechanism for the bit cell are two fold. First, this mechanism can introduce a threshold voltage offset in the SRAM devices with respect to isolated logic devices and second, it is an additional source of non-random mismatch and variation in channel doping.

2.4.4 Photo-resist implant shadowing

Because of the photo-resist thicknesses during the halo implant step, implant shadowing is another physical mechanism that becomes nearly unavoidable in the dense SRAM designs. The halo or pocket implant, used to control short channel effects, is commonly implanted at angles in the range of 30-45 degrees as a quad implant. Because of the pushed rules in the SRAM cell, the thickness and proximity of the photo-resist will result in some degree of implant shadowing in the dense SRAM devices. This has the potential of inducing threshold voltage shifts in the SRAM devices relative to the logic devices and for the type 4 topology, can also become a source of non-random mismatch. In addition to align-

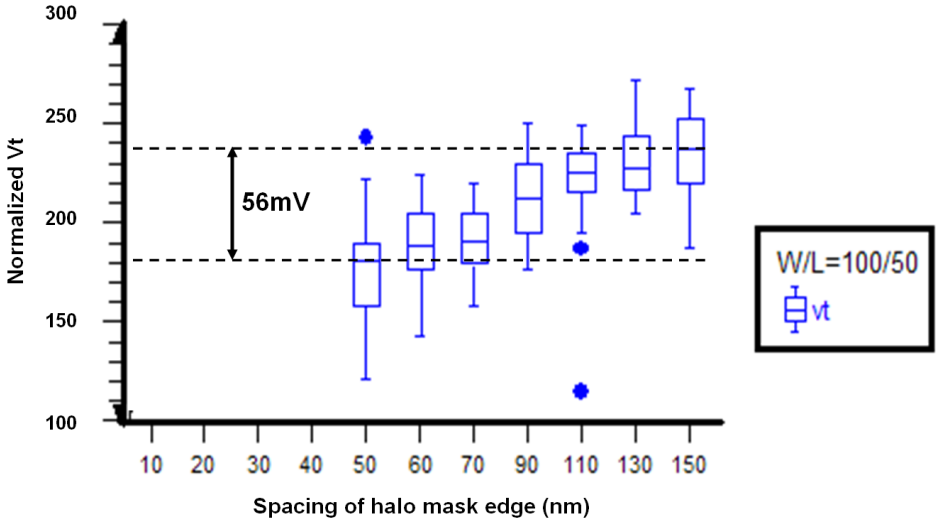


Figure 2.13: Measured hardware data showing effect of halo mask shadowing on narrow NMOS threshold voltage from 65nm process technology. (Mask edge is orthogonal to gate consistent with Fig.2.2.) Single points at 110nm and 50 are statistical outliers.

ment, this mechanism has the added variation components of resist thickness, and surface corner rounding.

With the resist thicknesses and design rules used in high density SRAM, halo shadowing, Fig. 2.8(d) will occur for at least one of the 4 quad implants for the PD and PG NMOS devices. TCAD simulations of a narrow 45nm NMOS indicate that this effect can be on the order of 50mV for a fully shadowed halo in the direction parallel with the polysilicon. Measured electrical results Fig. 2.13 are approximately consistent with the process simulation results and show a 56mV delta threshold voltage when the single quad halo (parallel with the gate) is fully blocked.

Due to the symmetry of the type 4 cell, halo block mask misalignment will result in a within-cell device threshold imbalance. As with the other alignment sensitive mechanisms previously discussed, use of a more aggressive n+ to p+ space will need to be compensated

Table 2.4: Dependencies and impacts of four mechanisms of non-random mismatch

Mechanism	**Primary dependencies**[a]	**Devices impacted**	**Topology 4 within cell mm**	**Topology 1x adjacent cell mm**
Implant straggle in SiO_2	Well dose, energy, species	PD,PU	$\mu_{Vtmm} \neq 0, \sigma_{Vt,DF} \uparrow$	$\mu_{Vtmm} \neq 0, \sigma_{Vt,DF} \uparrow$
Polysilicon inter-diffusion	Temp, dose, diffusivity	PU	$\mu_{Vtmm} \neq 0, \sigma_{Vt,GWF} \uparrow$	$\mu_{Vtmm} \neq 0, \sigma_{Vt,GWF} \uparrow$
Lateral ion straggle in resist	Well dose, energy, species	PD,PG,PU	$\mu_{Vtmm} \neq 0, \sigma_{Vt,DF} \uparrow$	$\mu_{Vtmm} \neq 0, \sigma_{Vt,DF} \uparrow$
Halo shadowing	resist thickness, halo dose, angle	PD,PG,PU	$\mu_{Vtmm} \neq 0, \sigma_{Vt,DF} \uparrow$	$\mu_{Vtmm} \neq 0, \sigma_{Vt,DF} \uparrow$

[a] dependencies in addition to n+ to p+ space, alignment, and CD variation

with improved alignment and/or image tolerance improvement to reduce the sensitivity to mismatch associated with this mechanism.

2.4.5 Mechanism impact summary

Table 2.4 summarizes the process dependencies, SRAM devices impacted, and compares the net effect differentiated by cell topology choice for each of the four mechanisms investigated. The net device impacts associated with the four mechanisms previously discussed will be dependent on the bit cell symmetry. Due to the symmetry differences between topology 4 and 1x, any deviation in alignment will translate directly to either a mean shift between device pairs within the cell or adjacent cell to cell device mismatch. The net measured effect of non-random mismatch is identified by the observation that ($\mu_{Vtmm} \neq 0$) for the local device pairs.

In addition to non-random mismatch, the standard deviation in Vt is also impacted by these mechanisms. The increase in standard deviation, as observed with polysilicon gate interdiffusion, can be large, Fig. 2.10(b), if the image tolerance, alignment, and additional processing dependencies such as temperature, grain boundary diffusion, and implant dose are not well controlled.

While all four systematic mismatch mechanisms identified are dependent on the n+ to p+ space, because type 1x cell area is limited by rules other than n+ to p+ space, the impact (assuming similar alignment and image tolerances) will be eliminated or significantly reduced for the 1x design relative to type 4. The amount of reduction will depend on the fabrication process details, in-line controls, the specific mechanism, and degree to which the n+ to p+ space can be optimally relaxed.

The density advantage of the type 4 topology would permit, if desired, optimized trade off of either increasing the n+ to p+ space or tightening the alignment and image tolerances of the implant blocking resist levels involved. Although the n+ to p+ space only occurs once in the Y_{1x} dimension, the 1x topology does not offer an intrinsic advantage in sensitivity to the n+ to p+ space. Taking the derivative of area with respect to (NP) for both topologies, equations (2.2) - (2.6), reveals that $2 \cdot Y_4$ is equal to X_{1x} when ($Lpu = Lpd$). If the n+ to p+ space became a limiting rule for the 1x topology, both topologies would be roughly equally impacted. However, because the rules for contacted gate to active area limit the 1x topology height, the n+ to p+ space is significantly more relaxed for this topology.

2.5 Non-random variation: Statistical infrastructure

A statistical infrastructure to establish the relationship between mismatch and margin limited yield is outlined in this section. To assess the impact of the non-random mismatch, the device threshold (Vt) will be treated as a continuous random variable. The 6T SRAM margin variance (σ_M^2) can be expressed as the sum of the squared components comprised of each of the 6 transistors shown in Figure 2.2. For reasons covered in the previous section, the SRAM device pairs may not have identical variances and therefore should be treated

independently for this analysis.

$$\sigma_M^2 = \sum_{i=1}^{6} \left(\frac{\partial M}{\partial Vt_i} \sigma_{Vti} \right)^2 \tag{2.8}$$

The margin value may refer to the read static noise margin (RSNM) or write margin (WM) for example. While it is commonly assumed that the population mean and variance of each of the 3 pairs of transistors in the cell are equal, deviations from this assumption can occur and are influenced by cell topology, process scaling and the used of pushed design rules. The margin mean is expressed in terms of (Vt) using the truncated form of the Taylor series expansion [68] as:

$$\mu_M \approx M_{Vt_{nom}} + \frac{1}{2} \sum_{i=1}^{6} \frac{\partial^2 M}{\partial Vt_i^2} \sigma_{Vti}^2 \tag{2.9}$$

The margin ($M_{Vt_{nom}}$) value at nominal Vt, refers to any margin which has a primary dependency on the device threshold. The margin limited yield may then be assessed by determining the fail probabilities. By accounting for the mean and variance components individually the fail probability may be computed from the standard normal probability distribution function (PDF). When no systematic or non-random Vt mismatch exists, the probability of failure (P_T) for the bit cell may be expressed as:

$$P_T[M \leq 0] = 1 - \int_0^\infty f_x(x)dx \approx erfc\left(\frac{\eta_{\sigma_M}}{\sqrt{2}} \right) \tag{2.10}$$

where η_{σ_M} is defined as $(\bar{M} - 0)/\sigma_M$. The probability is computed assuming a symmetrical 2-tail distribution to account for both states of the bit cell. When non-random Vt asymmetry exists, the fail probability for the left and right side of the cell must considered independently. This is expressed as:

$$P_T[M \leq 0] = P_R[M \leq 0] + P_L[M \leq 0] \tag{2.11}$$

The yield for a large array with (N_b) bits is then computed from the binomial relationship. Assuming no redundancy, the yield is given as:

$$Y_M = (1 - P_T[M \leq 0])^N b \tag{2.12}$$

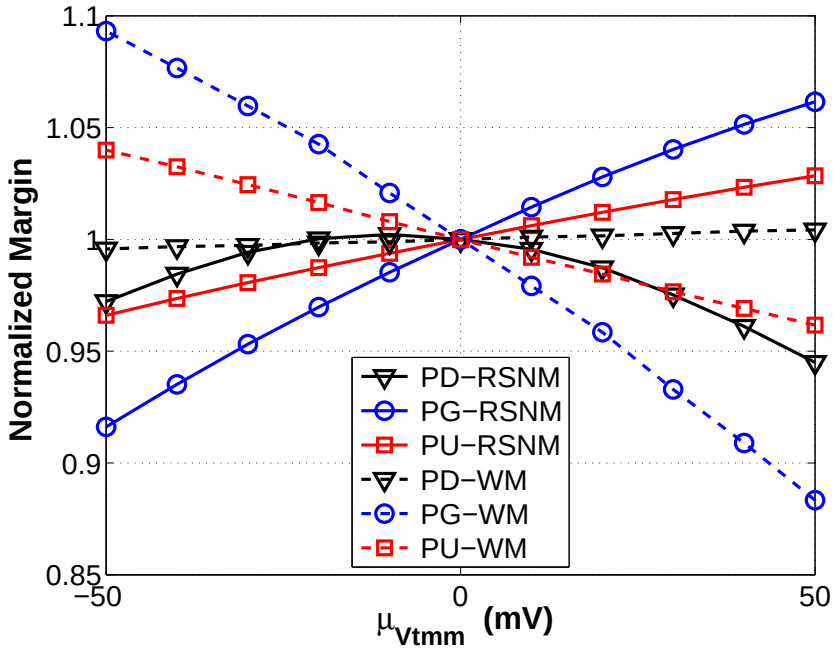
The margin mean and variance are explicitly dependent on the variance in device threshold voltage as described in (2.9) and (2.8). The more significant underlying components of the local variance in Vt are due to dopant fluctuations (DF), gate work function (GWF), and line edge roughness (LER). Treating these three components as independent random variables, the total variance is expressed as shown in (2.13).

$$\sigma^2_{Vt_{total}} = \sigma^2_{Vt,DF} + \sigma^2_{Vt,GWF} + \sigma^2_{Vt,LER} \tag{2.13}$$

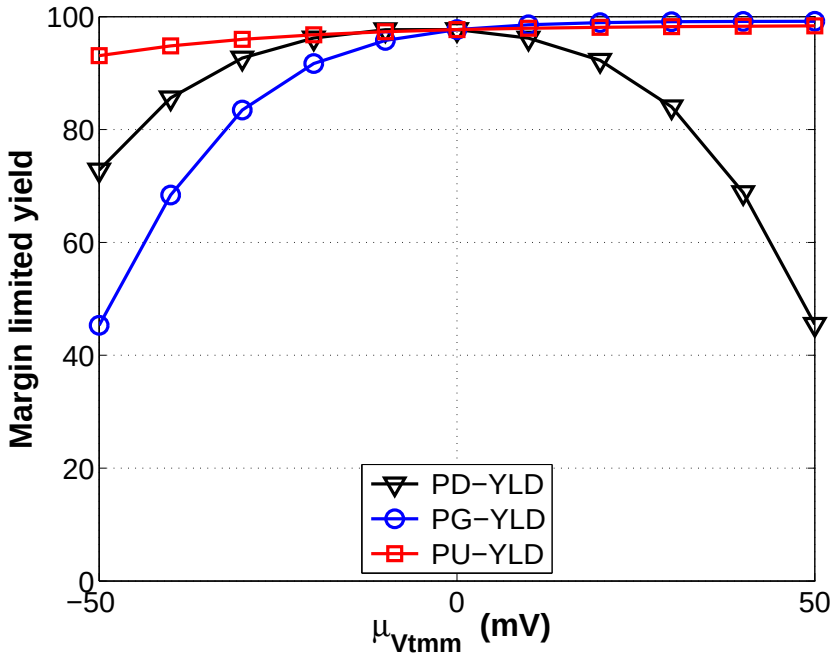
Processes or design topologies which are more susceptible to increases in the variance of any of these components are therefore less desirable. The dopant fluctuations (DF) are comprised of both random dopant fluctuations as well as any process induced systematic variations.

2.6 Quantifying the impact of non-random mismatch on yield

Random variation in device threshold is anticipated with an expected mean of zero for within cell mismatch. When non-random sources of mismatch produce a mean shift in



(a) Impact of systematic mismatch on RSNM and WM (normalized).



(b) Impact of systematic mismatch on margin limited yield (2 Megabit).

Figure 2.14: Impact of $\mu_{Vtmm} \neq 0$ on both RSNM and WM and margin limited yield. Simulations performed using on commercial 45nm LP technology SRAM models without the impact of increased variance.

$Vtmm$, such that ($\mu_{Vtmm} \neq 0$), an impact on the read static noise margin (RSNM), or write margin (WM) may be observed. To illustrate this, margin simulations are conducted using a commercial 45nm LP technology. The impact of μ_{Vtmm} on the mean RSNM and WM is plotted in Fig. 2.14(a). The margin limited yield, Fig. 2.14(b) for a 2 Megabit array is derived following (2.10), (2.11), and (2.12).

Given an equal amount of systematic mismatch, either within cell or adjacent cell-to-cell, a similar net yield impact is anticipated. This point is made by a consideration of the symmetries involved. For example, given the type 4 layout, the kinds of non-random mismatch described could result in the fail probability ($P[M \leq 0]$) for one of the two nodes becoming larger with respect to the other across the entire array. Alternately, for the type 1x symmetry, an adjacent cell-to-cell mismatch would be observed where the probability ($P[M \leq 0]$) for both nodes in every other cell in the array would be consistently increased. If both type 4 and 1x layout topologies were equally susceptible to the sources of non-random mismatch, the net impact on large array yield would therefore be negligible.

The effect of a non-random threshold shift (within cell or cell to cell), shown on the x-axis, is in addition to the background random variation that is present in the statistical models. For this technology, the PG NMOS exhibited the highest degree of margin sensitivity. At nominal voltage (1.1V) and room temperature, this bit cell is RSNM limited therefore the observed yield impacts, Fig. 2.14(b), are pronounced as the PG NMOS Vt is lowered. The RSNM and corresponding limited yield is adversely affected by both positive or negative shifts in the PD NMOS Vt. The PU PMOS has a more limited overall impact, lowering the RSNM limited yield as it becomes weaker.

2.6.1 Identifying non-random variation

The detection of alignment driven non-random offsets may prove to be difficult following typical manufacturing test restrictions. Although a large sample size may be accumulated or exist for the bulk population, alignment will vary for individual lots. The sample size required to detect a 10mV offset in the sample mean ($\widehat{\mu_{Vtmm}}$) with 95% confidence will be on the order of 60 or more as given by (2.14) which may exceed the number of samples tested for a given lot or batch of wafers run with a given alignment. If alignment and printed dimensions are centered and normally distributed, the entire population of Vt mismatch (right - left) will also be Gaussian in distribution. Systematic offsets in image critical dimension (CD) or alignment will impact the total population.

The population standard deviation can be derived from the unbiased estimator, $\widehat{\sigma_{Vtmm}}$, having a confidence interval as defined in (2.15). A method using Fourier analysis to separate the global variation sources from short range (mismatch) sources such as random dopant fluctuations (RDF) can also be applied to the SRAM devices [69]. To quantify the base line Avt expectation for a given technology [33], care is taken to eliminate channel proximity to drawn corners, well edges or block-level-resist edges. Characterizing $\widehat{\sigma_{Vtmm}}$ for identically drawn device pairs in close proximity across a range a channel areas provides the technology base line for NMOS and PMOS devices.

SRAM device pairs within a cell or cell to adjacent-cell are in close proximity and are drawn identically, however they are subject to additional sources of variation. These additional sources of dispersion in mismatch may be attributed to the pushed rules and layout topology used in the 6T bit cell. For advanced nanoscale technologies the threshold mismatch values measured for larger L dimensions are observed to fall on a different slope

than that of the minimum L [39] [10], therefore the Avt slope should be derived using the same L values as used in the cell.

Because the alignment sensitive occurrences of non-random mismatch can be limited to individual groups of samples, detection can be a challenge. A brief description of an approach for determining if such a condition exists is briefly summarized to illustrate a simple case where the sample variances can be assumed to be equal. Because the expected value of the paired data sample ($Vtr - Vtl$) mismatch mean is always zero, this is the null hypothesis. The confidence interval for testing this hypothesis is given by:

$$\mu_{Vtmm} \pm t_{\alpha/2} \left(\frac{\widehat{\sigma_{Vtmm}}}{\sqrt{N}} \right) \tag{2.14}$$

where N refers to the sample size, $\widehat{\sigma_{Vtmm}}$ is the sample standard deviation and $t_{\alpha/2}$ is the test with specific significance, α with ($N - 1$) degrees of freedom. The χ^2 distribution may be used to determine the confidence interval on the sample variance. The confidence interval may be expressed in the form of a probability, as:

$$P \left[\frac{(N-1)\widehat{\sigma^2_{Vtmm}}}{\chi^2_{N-1,1-\alpha/2}} \leq \sigma^2_{Vtmm} \leq \frac{(N-1)\widehat{\sigma^2_{Vtmm}}}{\chi^2_{N-1,\alpha/2}} \right] = 1 - \alpha \tag{2.15}$$

where the probability, P, (with significance level α) that the population variance lies within the defined intervals as defined by the χ^2 distribution.

2.7 Conclusions

Dopant fluctuations in nanoscale SRAM devices may be attributed to both random and non-random components. Cell layout topology, process scaling, and pushed design rules used in dense SRAM bit cell designs can influence the susceptibility to non-random mis-

match in present and future nanoscale SRAM devices. Four potential sources of non-random device mismatch that can impact dense SRAM designs were investigated. Two different bit cell topologies were considered to demonstrate how systematic mismatch decreases the margin limited yield in both topology types. For this reason, reduced dimensions required for the competitive SRAM bit cell as scaling continues below 32nm will place increased demands on the alignment and image tolerances required for large scale SRAM arrays.

Chapter 3

6T SRAM cell topologies for sub-22nm

3.1 Introduction

The extent to which the 6T bit cell can be extended through continued scaling is of enormous technological and economic importance. This chapter further addresses the challenging and complex cell design constraints being faced by the industry in CMOS process technology today and develops an alternative bit cell layout topology. Understanding the specific lithographic limitations and the mechanisms which drive systematic mismatch provides direction in identifying more optimum solutions. At the time of this work, the worlds' leading advanced silicon providers are developing and perhaps qualifying the 22nm (or 22nm/20nm) generation technologies, and work is underway on the 16/15nm node.

Beyond 22nm, it is certainly less clear if the planar 6T cell will maintain its dominant role in microprocessor cache, ASIC and mobile computing applications. This will depend on many factors such as continued advances in lithography, the successful incorporation of circuit assist methods [57], improved manufacturing practices for SRAM, and emerging

technology options to address variation. Additionally, the bit cell design may continue to evolve and adapt to the lithographic capabilities and constraints. In this chapter, the learning from chapter 2 is incorporated and expanded to develop the proposal of a new 6T bit cell topology for future nanoscale SRAM technologies.

The success of the (type 4) bit cell topology used in today's 6T SRAM is evident by its ubiquitous use in the advanced VLSI (65nm and below) technologies. Despite the widespread use of this bit cell, there are emerging challenges as scaling continues. The central question addressed in this chapter is, given the widespread use and acceptance of the type 4 topology as the optimal solution, "Do competitive 6T alternative topologies exist for 22nm and beyond, and if so, what might they be?"

3.2 Constraints and metrics for future nanoscale 6T bit cell

The desired attributes for the next generation bit cell topology would include high density, reduced lithographic and manufacturing complexity, low bit line capacitance, and elimination or reduction of the sources of systematic mismatch. At 22nm, the use of 193nm immersion lithography and double patterning will be employed by the leading advanced silicon providers to meet the aggressive layout dimensions required. For nodes below 22nm, extreme ultra violet (EUV) with a wavelength of 13.4nm will be phased in for the most critical layers. These changes may serve as a driving force for continued evolution of the 6T layout topology.

Reduced variation from lithographic sources, will continue to drive geometric simplic-

ity, pattern regularity, fixed pitch regulations and will continue to rely heavily on optical proximity correction (OPC) algorithms to meet the growing complexity [49] [38]. The use of double patterning is now commonly practiced for the gate level and pitch doubling techniques are being developed with renewed emphasis. Mask costs continue to increase with each new node, and the need for improved overlay or alignment tolerances drive increased costs of the stepper tools.

Printing the SRAM cell shared contact and conventional contact using the same mask level has been highly challenging and becomes more so as scaling continues. Elimination of right angles and jogs in the printed gate structures has been adopted for image control and integrity. Additional restrictions on gate direction and pitch are commonly implemented to provide further image fidelity. These factors converge to provide constraints on the cell designs for future technologies. These evolving constraints are becoming more restrictive with each node and effectively limit the viable set of 6T topologies for future nodes.

For the industry standard (type 4) 6T bit cell topology, there are several areas that are becoming more difficult with continued scaling. Two areas specifically highlighted are: 1) the metal 1 (M1) pattern required (Fig. 3.1) for the type 4 bit cell retains the relatively complex orthogonal directionality of the short lines [26], and 2) the jogs in the active silicon region, used to achieve a desired pull down to pass gate ratio for cell stability during a read access, are subject to significant rounding.

Given the growing lithography restrictions with scaling and the known 6T topologies [37], discussed in chapter 2, only two existing 6T topology options appear viable for further development. The topology that is currently the industry standard 6T cell (type 4) and a variant of the type 1 as covered in chapter 2.

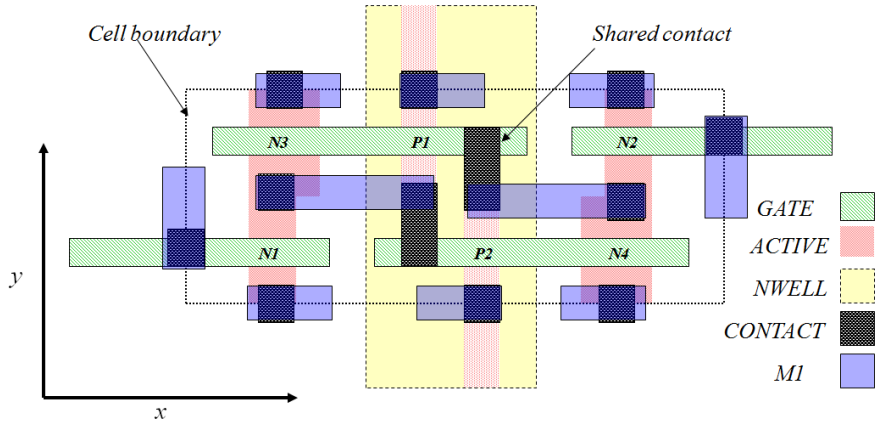


Figure 3.1: Type 4 6T layout (as shown in chapter 2, with the added drawn M1 layer. Depicts M1 layer pattern similar to that shown in reference [26], where the 'L' shaped pattern used in prior generations is eliminated to further simplify the required pattern.

In light of this, a re-examination of Ishida's four base layout categories may be useful to determine if additional suitable base category alternatives may exist. In this work, it is proposed that the four categories may be expanded to five as shown in Fig. 3.2. A new base category is achieved by shifting the placement of the cross coupled inverters so that the gate of the second inverter is in line with the contacts of the first inverter [59]. This new category provides a third viable 6T cell option, consistent with the deeply scaled CMOS lithographic restrictions and exhibiting many of the desired characteristics for further investigation.

The full 6T topology for this category 5 topology is shown in Fig. 3.3(a). There are potentially several advantages for future generation technologies with this new layout topology. First, the metal 1 (M1) complexity is reduced, further simplifying the required pattern compared the type 4 cell. Second, the cell height is further reduced (in the bit line direction) which allows for a reduced bit line capacitance and third, the jogs in the active silicon region are eliminated. This third point and its potential importance is explored in the next

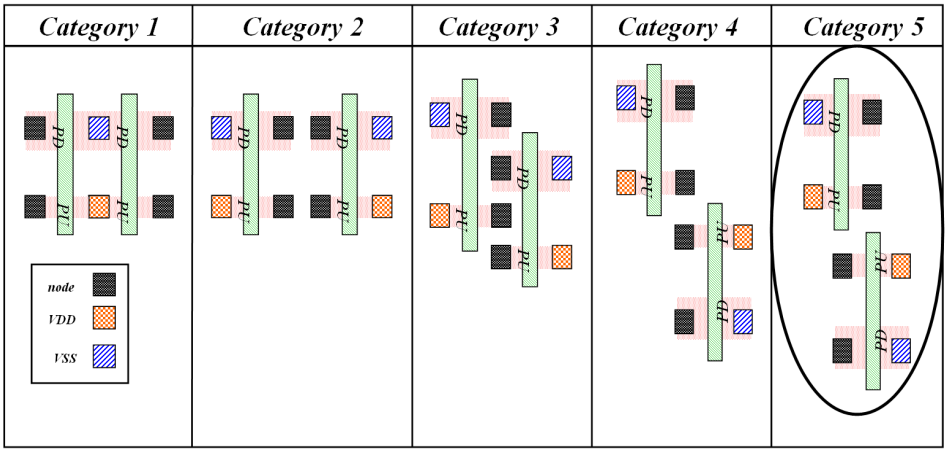


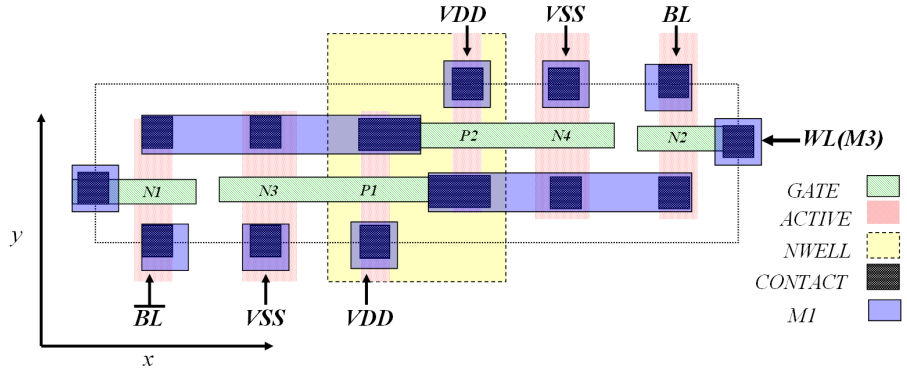
Figure 3.2: An additional category for the 6T layout is proposed. The cross coupled inverters are now shifted so that the gate of the second inverter is in line with the contacts of the first inverter.

section.

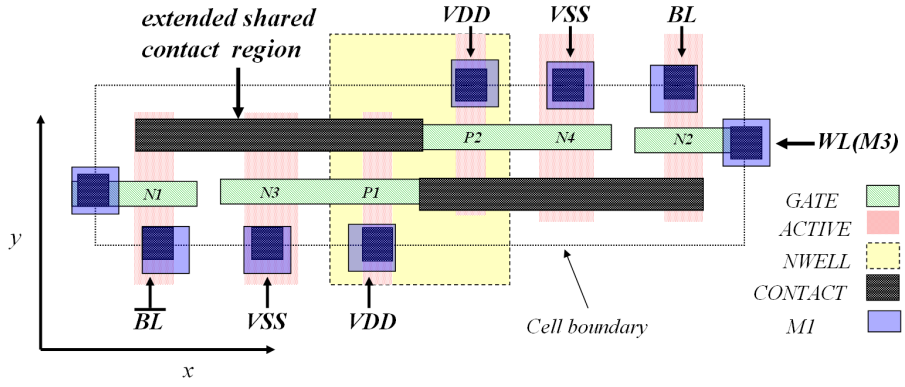
The shared contact is expected to have a similar complexity level as the type 4 where the shared contact 'bar' and conventional contact 'square' regions are printed using the same mask. An alternative layout scheme extends the shared contact across diffusion regions of the opposite inverter. This alternative layout Fig. 3.3(b) , requiring an elongated 'shared contact' may offer unique options that will be discussed in more detail later.

3.2.1 Additional sources of device variation in SRAM

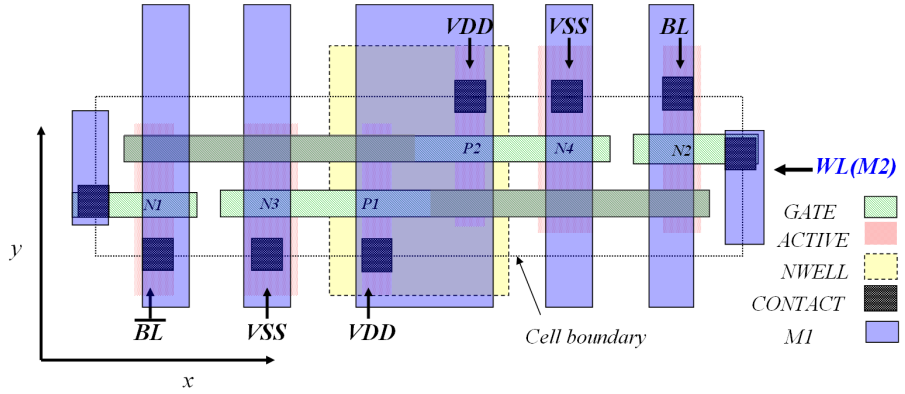
Sources of non-random mismatch associated with dense nanoscale SRAM devices discussed in chapter 2 specifically addressed the systematic sources of mismatch which were due to variations in channel doping (both random and systematic). The general subject of non-random variation in dense SRAM devices may be further expanded to include the geometric sources of mismatch. These arise from the non-ideal environment associated



(a) Version 1 of new category 5 6T bit cell topology (type 5).



(b) Version 2 of new category 5 6T bit cell topology with extended shared contact(type 5e).



(c) Version 3 of new category 5 6T bit cell topology with replacement gate and buried contact(type 5b).

Figure 3.3: Various layout options for new category of ultra-thin (UT) 6T bit cell topology with reduced M1 lithography complexity, reduced bitline capacitance, and reduced mismatch due to corner rounding in the active silicon.

with pushed design rules, variation in alignment and additional lithography effects such as corner rounding and line end foreshortening. These effects are layout topology dependent and can also contribute to the overall mismatch in the dense bit cell devices. Accounting for these additional components, the total variance is then expressed more fully as:

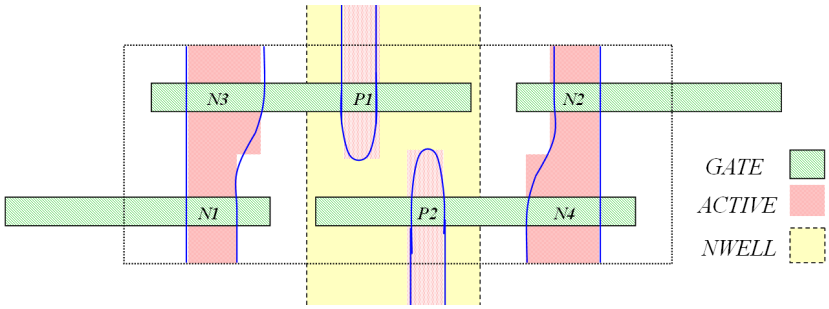
$$\sigma^2_{Vt_{total}} = \sigma^2_{Vt,DF} + \sigma^2_{Vt,GWF} + \sigma^2_{Vt,LER} + \sigma^2_{Vt,Weff} + \sigma^2_{Vt,Leff} \tag{3.1}$$

where the first term, $\sigma^2_{Vt,DF}$, captures the variation in channel doping due to both random and sources of systematic variation described previously. The second term, $\sigma^2_{Vt,GWF}$ captures the variation associated with the gate work function. The last three terms in (3.1) capture the physical or geometrical variation. While line edge roughness (LER) plays a role in the ideal logic mismatch, the last two terms are typically neglected due to the proximity assumptions of the drawn ideal mismatch structures. As illustrated in Fig. 3.4, this is not always the case for the dense SRAM devices.

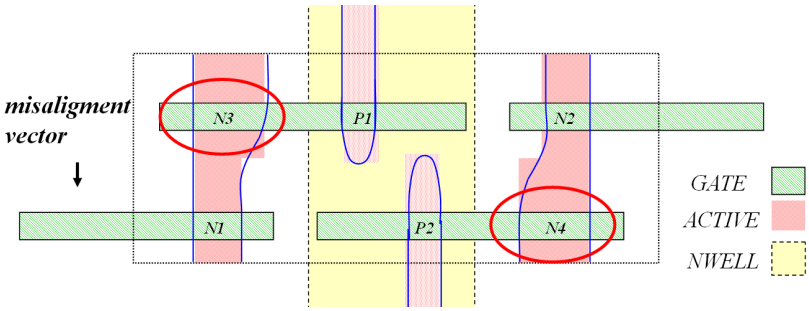
The geometry of the right(N4) and left(N3) devices, Fig. 3.4, become increasingly dissimilar as a function of alignment. Additional variation in the Leff (not shown) can arise from similar arguments when line end foreshortening coupled with corner rounding are captured for this layer. Although alternate notchless cell options have been proposed previously [92] [42] [43], to avoid this form of within-cell mismatch, they have not been adopted by the industry.

3.2.2 Estimation of the new 6T bit cell area

By using the set of pushed layout rules, given in chapter 2, optimized for the type 4 layout, the bit cell area for this topology may be estimated for comparison purposes.



(a) Active region corner rounding illustrated (solid lines outline of active region) with nominal gate to active alignment.



(b) With misalignment the PD NMOS devices become geometrically mismatched due to corner rounding effects associated with the jog.

Figure 3.4: Illustration showing impact of gate misalignment on the device geometries. The devices circled exhibit different width characteristics and the width of N3 is effectively less than that of N4.

Following the pushed scaling rules defined in chapter 2, the X_5 dimension is estimated to be approximately:

$$X_5 = 2 \cdot (\frac{1}{2}(CW) + (GCA) + (Wpg) + (GPA) + (TT) + (GPA) + (Wpd) + (NP) + (Wpu) + \frac{1}{2}(AA)) \tag{3.2}$$

and the dimension (Y_5) is calculated to be:

$$Y_5 = 2(\frac{1}{2}(CW) + (GC)) - (\frac{1}{2}((CW) + CS + CW) - (CW - max(Lpd, Lpu)))) \tag{3.3}$$

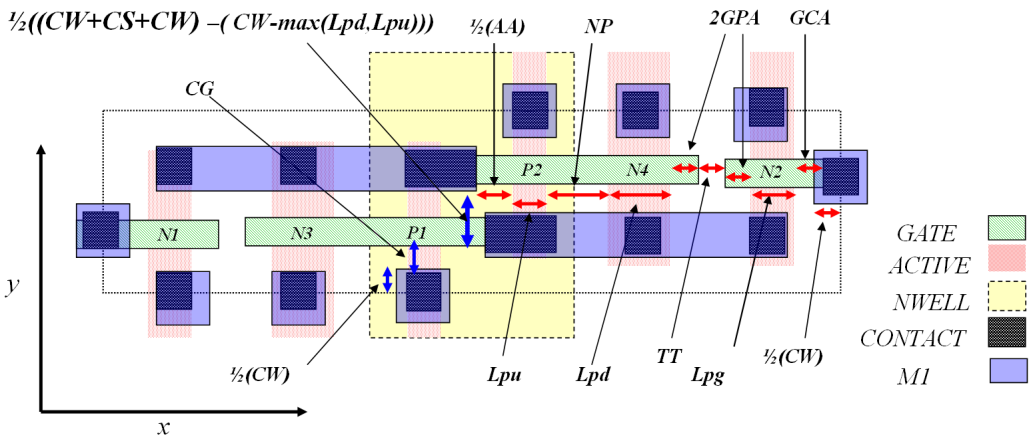


Figure 3.5: Type 5, 6T layout with the area limiting rule assumptions highlighted.

Using consistent assumptions the Y_5 estimate of 6.5λ represents more than a 13% reduction in the bitline length over the array. This directly corresponds to improved access speed. The cell area is therefore:

$$A_5 = 6.5\lambda \cdot (14.7\lambda + 2\lambda(Wpu + Wpg + Wpd)) \tag{3.4}$$

Using scaled and equivalent device dimensions a comparison of the calculated bit cell area results in $168.35\lambda^2$ for cell type 5, and $142.45\lambda^2$ for type 5e, compared to the $120\lambda^2$ estimated for the type 4 cell. The limiting design rules used to calculate the type 5 cell dimensions are highlighted in Fig. 3.5.

The second layout method which utilizes the extended shared contact (referred to as type 5e) is used to illustrate the potential area improvement that could be obtained by using a pitch doubling technology. While the assumed X_5 value will remain equivalent to the the type 5, the Y_5 value could be further reduced by:

$$Y_{5e} = 2(\tfrac{1}{2}(CW) + (GC) + max(Lpd, Lpu, Lpg) + \tfrac{1}{2}(GS)) \tag{3.5}$$

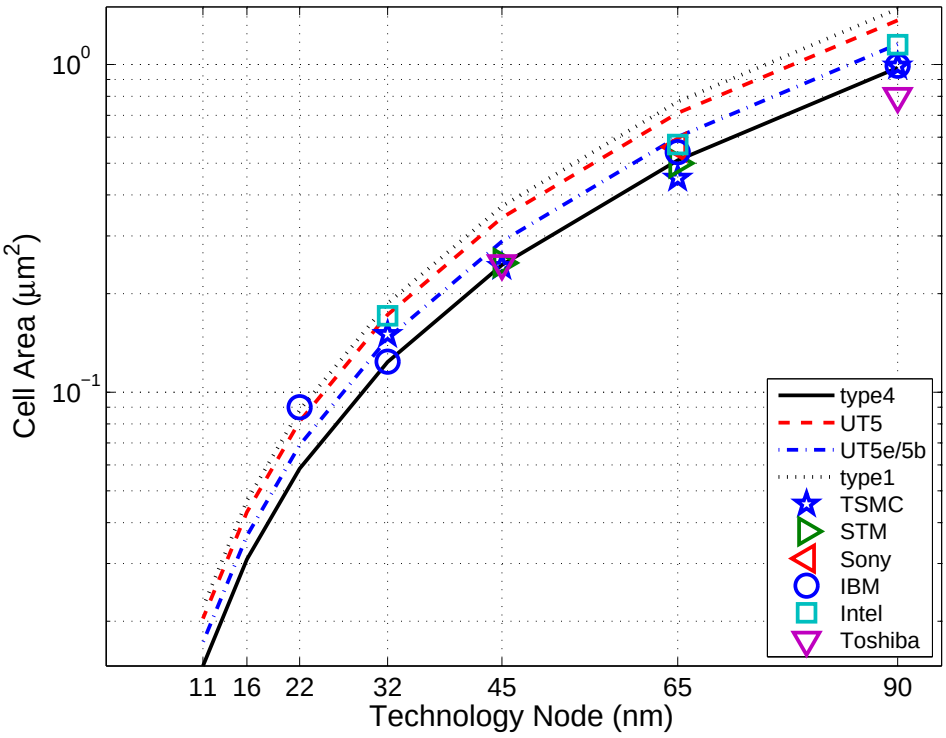


Figure 3.6: Calculated area for topology 5 cell across multiple technology nodes.

The area for the type 5e becomes:

$$A_{5e} = 5.5\lambda \cdot (14.7\lambda + 2\lambda(Wpu + Wpg + Wpd)) \tag{3.6}$$

The type 5b example demonstrates the potential for further development. The potential synergy with the replacement gate process option used by some to combine the gate and shared contact patterning is clearly and interesting possibility for further exploration. By replacing the three 'in-line' shared contacts shown in Fig. 3.3(c) with shared buried contacts (patterning the gate and shared contact in one step) is an area for further investigation. If the shared contact layer is separate and isolated from the conventional contact, the cell can be wired very simply with VDD, VSS,BL and BLB running vertically (y-direction) using the M1 layer and with the WL on M2 completing the 6T design running horizontally. This

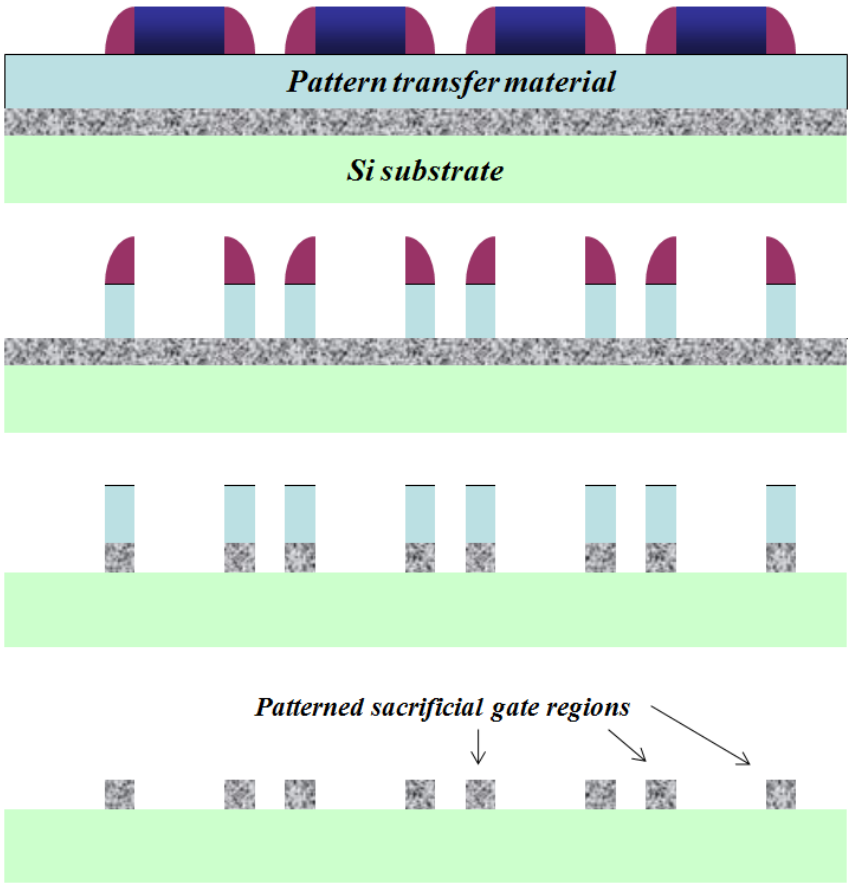


Figure 3.7: Cross section view of gate pattern method where array is first patterned by a series of continuous lines using sidewall image transfer technology.

is an area for future exploration since the advantage of reduced required metal levels to complete the cell could offer savings in cost and free up M3 wiring channels across the array for logic routing.

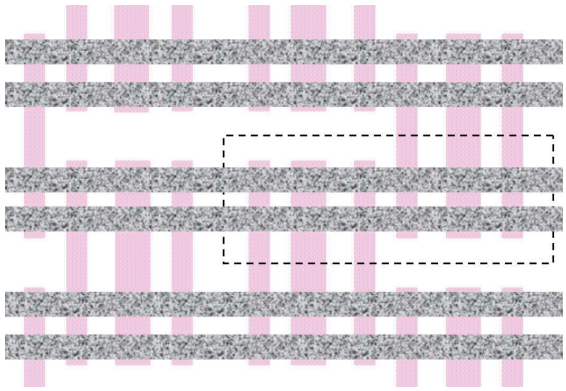
A process flow and full array layout is next developed to further demonstrate the potential advantages of this new cell topology. Optimum gate spacings may be obtained using a pitch doubling technique which also provides a means of achieving the optimum line spacings. A cross section process flow for this is given in Fig. 3.7.

A top down view showing the active regions in an array segment, with the continuous lines formed following the flow in Fig. 3.7 is shown in Fig. 3.8(a). The cut mask is then used as shown in Fig. 3.8(b) to complete the gate pattern. The final gate pattern is then created as shown in Fig. 3.8(c). A dashed rectangle outlining the boundary of a single bit cell is replicated for continuity across Fig. 3.8, Fig. 3.9 and Fig.3.10.

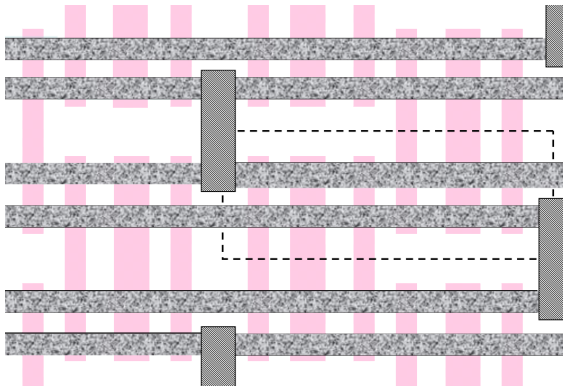
An insulating dielectric is then applied followed by a chemical mechanical polish (CMP) to planarize the surface and expose the top of the gate pattern sacrificial material, Fig. 3.9(a). The original gate sacrificial material is then removed and the gate dielectric material (e.g., high-κ) is deposited. A buried contact mask is used to remove the gate dielectric in regions where the buried contact is desired. The gate material is then deposited followed by a CMP to planarize and self-align the gate material and gate dielectric to the predefined openings, Fig. 3.9(b). An additional insulating dielectric is deposited and the contact mask, etch, deposition and CMP is used in the conventional way to form the contacts as shown in Fig. 3.9(c).

To complete the wiring, which requires only two metal levels, the M1 lines may be either printed as continuous lines, followed by a cut mask or printed with a single mask depending on the specific constraints and tolerance requirements. All M2 lines are unidirectional with this cell layout topology as shown in Fig. 3.10(a). Only one M2 to M1 via per cell is required, Fig. 3.10(b), to allow the word line connection at M2. The M2 lines follow a regular unidirectional pattern as shown in Fig. 3.10(c). The array required M2 line/space will be much more relaxed than the typical minimum M2 pitch and and may permit additional wiring tracks for global signals or logic as needed.

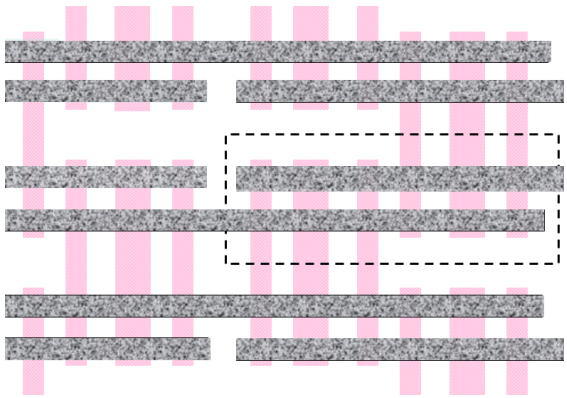
Using a consistent set of pushed SRAM layout rules, the newly defined bit cell does



(a) After processing shown in Fig. 3.7.

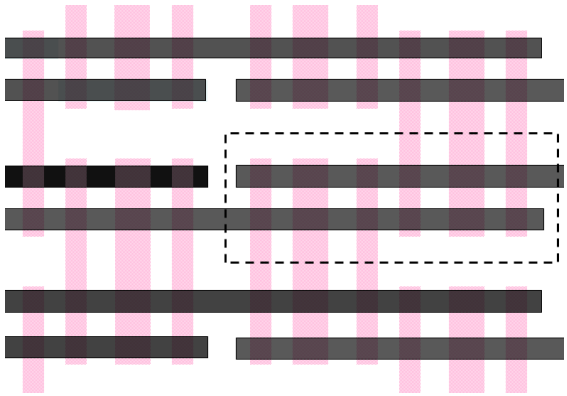


(b) Gate cut mask.

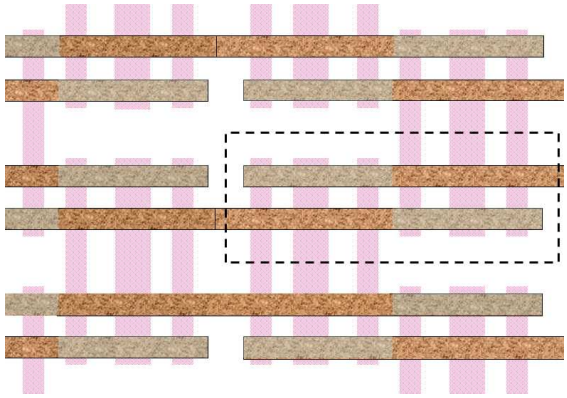


(c) Final gate pattern.

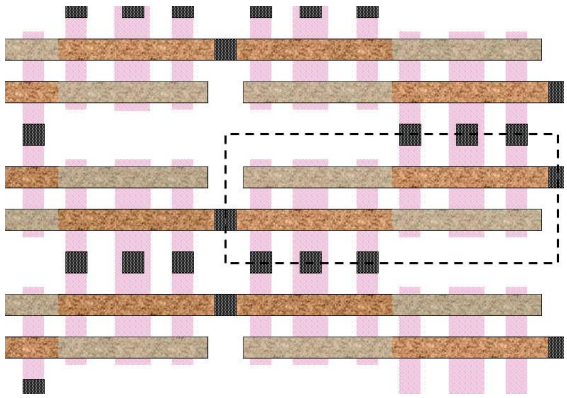
Figure 3.8: Top view of an array gate segment showing patterned active silicon regions and the gate definition sequence. (a) Continuous gate lines running horizontally (following processing shown in Fig. 3.7). (b) Dual pattern gate cut mask indicating openings in resist to allow completion of the gate pattern. (c) Following dual pattern cut mask processing, the final gate pattern is completed. The dashed rectangular region outlines area of a single bit cell.



(a) After CMP and sacrificial gate material removal.

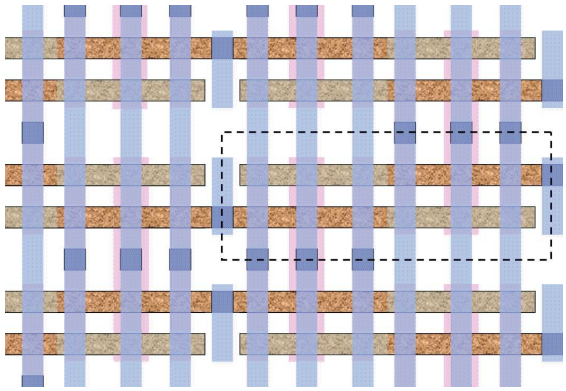


(b) Gate deposition and CMP.

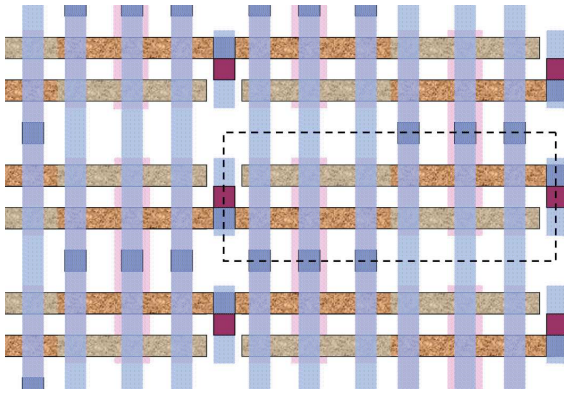


(c) Conventional contact formation.

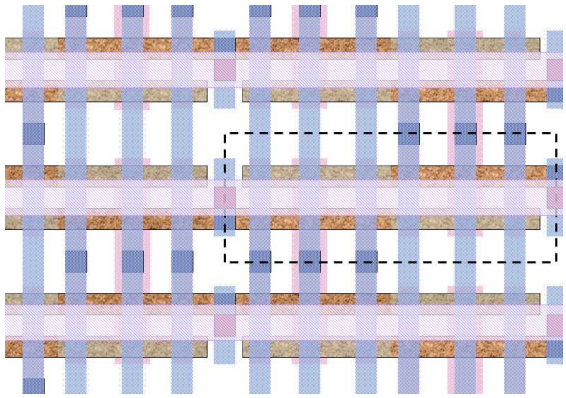
Figure 3.9: Top view showing array buried contact and final gate processing sequence. (a) Top view of array segment showing areas where the gate sacrificial material was removed. (b) After dielectric deposition, buried contact mask processing, gate deposition and CMP. (c) Array segment after conventional contact formation steps. The dashed rectangular region outlines area of a single bit cell for continuity with the previous figure.



(a) M1 pattern (unidirectional vertical lines).



(b) Through V1.



(c) M2 continuous unidirectional horizontal lines.

Figure 3.10: Top view of array segment showing M1 through M2 patterned regions. (a) Top view showing M1 pattern of unidirectional lines running vertically. (b) Top view of V1. Only one via per cell is required for this cell topology. (c) Top view of M2 lines running horizontally. The dashed rectangular region outlines area of a single bit cell for continuity with previous figures.

Table 3.1: SRAM cell metric comparison

	cell type			
Metric	**4**	**5**	**5e**	**5b**
Number contacts	6	8	4	4
Number shared contacts	2	2	2e[a]	2b[b]
Cell area (λ^2)	120	168	142	142
L_{BL} (λ)	7.5	6.5	5.5	5.5
Number Metal levels	3	3	3	2

[a] designation 'e' refers to extended shared contact)

[b] designation 'b' refers to extended shared/buried contact)

not achieve the density calculated for the type 4 layout. This is partially due to the fact that the pushed rules used today are clearly optimized for the type 4 layout topology. The calculated cell areas based on the equations given here and published 6T bit cell areas are shown in Fig. 3.6. It is not clear if the deviation from 120 λ^2 is driven purely by W and L up-sizing or if lithography limitations are playing a larger role. This deviation from the traditional scaled area may indicate that the type 4 layout is hitting limitations in scaling which will renew interest in alternative topologies such as proposed here.

A comparison of bit cell metrics which highlight the key differences by cell type is given in Table 3.1.

A comparison of bit cell metrics which highlight the key differences by cell type is given in Table 3.1. The bit cell area, BL length (L_{BL}), and number of required metal levels is summarized. Because the number of contacts required per cell is also a metric of interest,

this metric highlights an additional advantage of the type 5 topology.

The new 6T cell layout allows active silicon regions, gate, M1 and M2 to be printed as a series of straight unidirectional lines across the array, as shown in Fig. 3.10 for a fully wired type 5e drawn array segment, eliminating the need for complex shapes corners and jogs. Active silicon, gate and M1 may be completed with a cut mask layer. Reduced systematic mismatch in the pull down NMOS devices as a result of the elimination of jogs in the active silicon. An improvement of 13% or more in read access delay may be realized due to reduction in the bit line length.

3.3 Conclusions

As layout dimensions continue to be reduced, lithographic considerations will impose additional constraints on the layout of future nanoscale SRAM layout. Previously identified 6T bit cell topologies offer few alternatives for further exploration beyond 22nm. A new 6T topology is proposed in this work which may offer improved compatibility with future lithography restrictions and provide some additional advantages over the existing type 4 topology. Based on this analysis, an area penalty of approximately ($18 - 40\%$) will need to be weighed against the advantages of reduced alignment sensitive geometric mismatch, improved performance through reduced BL capacitance and reduced lithographic complexity.

Chapter 4

Coping with variability: Circuit Assist Methods

4.1 Introduction

Large scale 6T SRAM beyond 65nm will increasingly rely on assist methods to overcome the functional limitations associated with scaling and the inherent read stability/write margin trade off. The primary focus of the circuit assist methods has been improved read or write margin with less attention given to the the implications for performance. In this chapter margin sensitivity and margin/delay analysis tools are introduced for assessing the functional effectiveness of the bias based assist methods and show the direct implications on voltage sensitive yield. A margin/delay analysis of bias based circuit assist methods is presented, highlighting the assist impact on the functional metrics, margin and performance. A means of categorizing the assist methods is developed to provide a first order understanding of the underlying mechanisms. The analysis spans four generations of low

power technologies to show the trends and long term effectiveness of the circuit assist techniques in future low power bulk technologies.

The 6T SRAM cell design has been successfully scaled in both bulk and SOI down to the 32/28nm node and has remained for more than a decade the dominant technology development vehicle for advanced CMOS technologies. Reduced device dimensions and operating voltages that accompany technology scaling have led to increased design challenges with each successive technology node. Large scale 6T SRAM beyond 65nm will increasingly rely on assist methods to overcome the functional limitations imposed by scaling and the inherent read stability/write margin trade off. An objective metric based methodology is developed for the evaluation of scaled CMOS technologies to provide guidance in the selection of assist methods. This chapter explores various assist options given the technological constraints, functional boundary conditions and scaling trends that must be addressed for successful migration beyond 32nm.

4.2 Background and Motivation

A unique feature of the 6T SRAM is an inherent balance between stability when holding data during a read or non-column selected write access and the ability of the cell to be written. This fact means that the device dimensions and threshold voltage targets established for the SRAM devices are a compromise by design. The ability to read and write will be characterized in terms of margins to assess the functional implications. These margins, will be referred to as write margin (WM), and read static noise margin (RSNM) or static noise margin (SNM), tend to decrease with scaling. When this fact is considered in context with the growth in bit count and increased variability with each successive generation, we may

better comprehend the true nature of the mounting concern. This work seeks to explore the circuit options that may be needed to overcome the collapsing window of functionality and to provide a methodology for evaluation of the circuit assist options.

With continued scaling, circuit assist techniques will become necessary to preserve the 6T cell functional window of operation as scaling continues. A variety of SRAM functional assist methods have been proposed, however there remains no clear agreement in the industry as to which method or combination of methods will emerge as the more optimal solution. Moreover, different works compare the assist features in varied settings of technology node and technology type, but little detail is given on the trade offs involved in the selection process. Therefore, one goal of this chapter is to develop an objective, metric based methodology to provide guidance for selecting an optimum assist feature for a technology platform. A second objective is to explore the impact of CMOS scaling trends on the robustness of various assist methods.

Circuit assist techniques will become increasingly necessary to preserve the 6T cell functional window of operation as scaling continues. A range of SRAM functional assist methods have been proposed and discussed, however there remains no clear agreement in the industry as to which method or combination of methods will emerge as the more optimal solution. While different works compare the assist features in varied settings of technology node and technology type, often little detail is given on the trade offs involved in the selection process. Although power and cost are clearly important factors in determining the optimal assist method, it is first necessary to determine if an assist method will meet the functional margin and delay requirements. Once the assist methods that meet the functional requirements are established, the power and implementation costs can by weighed. The

goal of this work is to provide an approach for assessing the functional effectiveness of the assist methods. A second objective is to explore the impact of CMOS scaling trends on the robustness of various assist methods. The specific contributions of this chapter include:

• A margin/delay analysis method is developed for the evaluation of the functional effectiveness of circuit assist methods in 6T SRAM.

• A concurrent analysis across four technology nodes to explore the potential impacts of scaling in low power bulk CMOS technologies.

• A concise overview, and method for categorizing the 6T SRAM assist options.

4.3 Assist categories

A categorization of the assist methods is introduced to establish a systematic means of characterizing the range of circuit assist techniques used in this discussion. For a given foundry cell design, there are three distinct circuit types or categories to address the reduced window of functionality for the 6T SRAM:

1. Altering noise source amplitude or duration through the access transistor,

2. Modification of the latch strength or voltage transfer characteristics of the latch inverters,

3. Avoidance or masking by design or architecture methods.

While category 3 is included for thoroughness and encompasses a range of approaches including ECC masking or prohibiting the half select issue during a write operation [16], the scope of this work will focus on the bias based methods as defined by type 1 and 2. A

categorized summary of the bias based circuit assist methods is shown in Table 4.1. The assist type given in Table 4.1 provides the primary mechanistic explanation for the assist method effectiveness. While the category types are useful for quickly analyzing the various assist techniques, they are not fundamentally exclusive, and in some cases both mechanisms influence the net assist effectiveness as will be discussed in more detail in section 4.7.

The read and write assist methods listed in Table 4.1 can and in many cases are used in combination, and most can be implemented in either a static or dynamic mode. The categories can be further distinguished by the voltage terminal or terminals which are manipulated. For example a change in the WL voltage would involve modifying one voltage level while a change in the global VDD would involve changing the voltage on 5 of the 7 available terminals associated with the 6T SRAM cell (VDDc, NWELL, WL, BL and BLB). Increased global VDD is unique for several reasons and will be discussed in more detail in section 4.6. Modification of the cell design parameters such as device W,L, or device threshold voltage by process change or by means beyond the control of the circuit designer, are outside the scope of this body of work.

4.4 Review of assist methods

A brief overview of circuit assist methods published over the last five years will support the objectives of this chapter, but the large number of publications prevents an exhaustive review here. It is sufficient for this purpose to provide a sample of the options that have been proposed and to allow us to discuss some of the major advantages and disadvantages in context of the categories and terminal access options given in Table 4.1.

Table 4.1: Summary of SRAM circuit assist methods with predominant assist type

Read Assist	**Type**	**Write Assist**	**Type**	**Terminal(s)**
Raise VDD	2	Raise VDD	1	global [a]
Raise VDD at cell	2	Reduce VDD at cell	2	VDDc
Reduce VSS at cell	2	Raise VSS at cell	2	VSSc
WL droop	1	WL boost	1	WL
Reduce Q on BLs [b]	1	Increase (BL-BLB)	1	BL & or BLB
Weaken pass gate NMOS	1	Strengthen pass gate NMOS	1	array PWELL [c]
Strengthen pull-up PMOS	2	Weaken pull-up PMOS	2	array NWELL

[a] VDD applied to terminals VDDc, WL, NWELL, (BL and BLB for read, BL or BLB for write)

[b] reduced voltage or capacitance on BL

[c] Well bias also modulates pull-down NMOS device in most bulk technologies

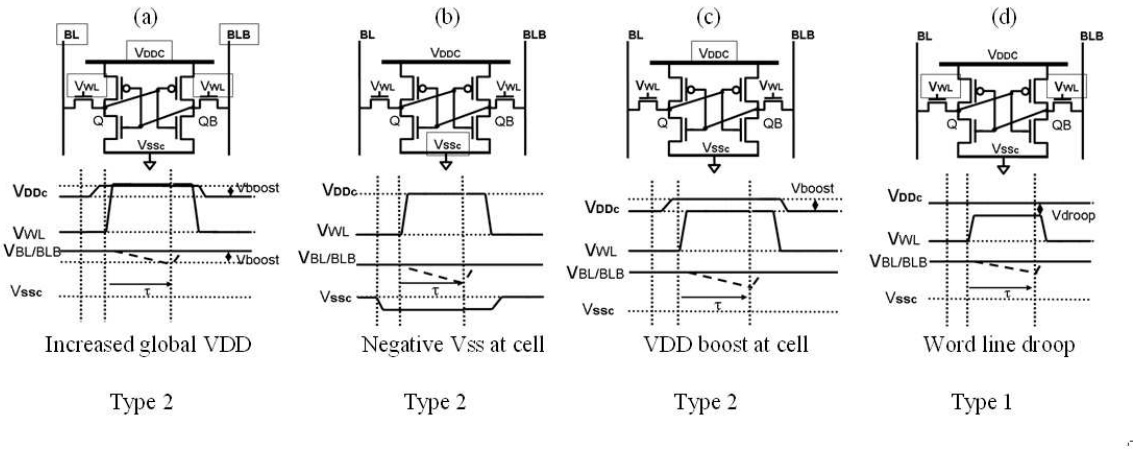


Figure 4.1: Schematic timing diagram representations for read assist (a) raised array global VDD, (b) negative VSS at the cell, (c) VDD boost at the cell and (d) WL droop. τ represents the time for the sense amplifier to set. Text box denotes modulated terminal(s).

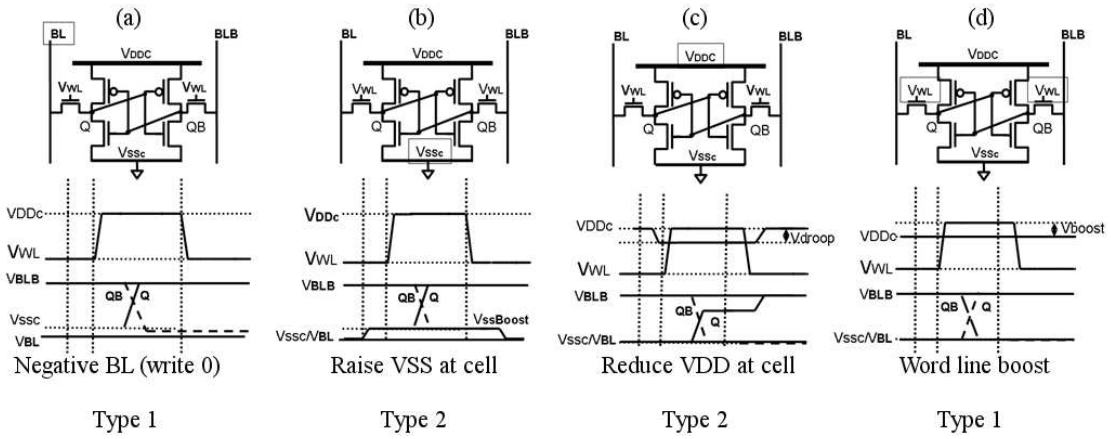


Figure 4.2: Schematic representations for write assist (a) negative BL, (b) raised VSS at the cell, (c) VDD droop at the cell and (d) WL boost. Text box denotes modulated terminal(s). Node voltage Q represented by dashed line in schematic timing diagram.

4.4.1 Read Assist

Those read assist methods categorized as type 1 include methods that reduce the noise source amplitude or duration, which impact the storage latch. These include the method of write-back [70][44][46], reduced word line gate voltage [66][29][84][61][64], increased word line (pass gate) threshold voltage through body bias [62][91], and reduced bit line charge by lowering the voltage or capacitance [44][8][9][1]. The methods categorized as type 2, which are intended to improve the resilience of the latch, are increased array VDD [29][18][19][95][92], decreased array VSS [84]and reduction in the absolute value of the SRAM pull up PMOS threshold voltage [62]. While some techniques such as write-back (or read-modify-write) are purely dynamic in nature, those techniques which involve altering the well (NWELL or PWELL) bias are proposed as primarily static implementations due to the large RC delay or layout complexity that would be involved in making this technique dynamic. The embodiments proposed as assists in [62][90] are essentially fixed biases set at one point in time to provide some compensation for global variation.

4.4.2 Write Assist

A roughly equal number of publications are invested in the challenge associated with writing the 6T SRAM. The read/write assist symmetry observed from table 4.1 is worth noting, and all but one method (increased global VDD) have the not so surprising opposite effect on read stability versus ability to write. Publications that address the challenge of writing the cell following category 1 (increased amplitude or duration of the write signal through the pass gate device) have proposed some form of boost to the word line gate voltage [29][18][35][19] or negative bit line voltage [84][77][64] to increase the VGS of the pass gate device. Those publications that address improving write margin by means of reducing the latch strength include reducing the array supply voltage VDDc [22] [70][66][29][61][91][95], raising the array VSSc [8][94][79], or reducing the strength the pull up PMOS by NWELL bias [62][90].

4.5 Assist Metrics

The primary objective of this work will be focused on the functional metrics of margin sensitivity and performance. The metrics of power and cost will be addressed in section 4.7 in context of this primary objective. In this section we define and quantify of the margin and performance metrics used in this analysis.

4.5.1 Margin Sensitivity

The margin sensitivity is defined as the change in margin with respect to the change in applied assist bias voltage for a given technique. This is expressed as:

$$Sensitivity = \frac{\partial(Margin)}{\partial V} \tag{4.1}$$

Margin may refer in this case to either SNM or WM. To compare the margin sensitivity of the specific assist methods, noise margin analysis is performed using custom predictive technology models (PTMs)[97][12] using pre-defined scaled SRAM dimensions consistent with the dense SRAM published values. The defined margin sensitivity is a useful metric for quantitatively comparing assist method effectiveness. It is applicable to all bias based assist methods, provides an objective means of comparing the assist methods to one another and also across the technology nodes.

Because bias limitations of some form exist for all assist methods, the margin sensitivity provides a means of quantitatively determining the attainable margin improvement. Depending on the assist method used, different limiting factors will constrain the terminal bias values that can be applied. In the case of boosting schemes such as +WL(write), neg BL(write), +VDDc(read) or -VSSc(read), the common limiting factor is the technology Vmax. Voltage suppression schemes such as +VSSc(write), -VDDc(write) or -WL(read) are limited by different mechanisms. For example, the bias used collapsing the supply voltage (+VSSc or -VDDc), becomes limited by data retention fails for unaccessed cells that share the collapsed supply. For -WL(read), performance limitations can quickly limit the allowable bias available for read stability margin gains obtained with reduced word line voltage.

The nominal VDD is based on published industry values for the nodes of interest. The

Vdd values used were 1.2V, 1.1V, 1.1V and 1.0V for 65nm, 45nm, 32nm, and 22nm respectively. As part of the methodology defined in this investigation, particular emphasis is placed on the specific conditions that represent the worst case operation voltage (Vwc) for the technology. Vwc is defined as the minimum voltage at which the SRAM must be able to perform both a read and write operation across the entire array without failure. Thus, one must ensure that the VDDmin[1] for a given array is at or below our predefined Vwc for each technology node. Because Vwc is recognized as technology and application dependent, 0.8X the nominal VDD will be used as this value. This condition accounts for factors such as voltage droop, NBTI shifts over the product lifetime, and testing equipment variability.

In addition to the shift in the mean margin value, variation and the impact of the assist methods on the margin distribution is also examined in section 4.6. This is a critical point since the ultimate goal of the assist technique is to improve the yield at the Vwc or lower the VDDmin of a particular array.

4.5.2 Performance

The performance for a given assist method is evaluated using write delay for the write assist method and the time required for bit line signal development for read assist. For this analysis, a concern about the deltas between techniques exists. This simplifies the analysis and allows us to focus specifically on the two performance components of interest. The delay can be reduced to the time required to charge the word line (τ_{WL}), plus the time required to develop a sufficient differential voltage on the BL ($\tau_{\Delta BL}$) to set the sense

[1]While non-foundry or in-house designs may have the flexibility to push the operation voltage to the empirically defined VDDmin, foundry based design kits frequently specify a valid model operation voltage range. Designing outside this specified range (below Vwc) may produce invalid results.

amplifier.

$$\tau_{read} = \tau_{WL} + \tau_{\Delta BL} \tag{4.2}$$

To briefly illustrate how the assist method may impact the τ_{Read}, the read assist method of reduced WL voltage is considered. For this example, the τ_{WL} will be reduced, while the $\tau_{\Delta BL}$ will be increased.

Following a similar approach as with the read performance evaluation, considering the deltas associated with the assist methods for comparison purposes, the write performance (τ_{write}) estimate will require three components as given in (4.3).

$$\tau_{write} = \tau_{BL} + \tau_{wcell} + \tau_{WL} \tag{4.3}$$

The value τ_{WL} is consistent with the previous definition, and τ_{BL} is the delay (or part of the delay that does not overlap with τ_{WL}) required to establish the BL-BLB voltage differential for the write operation. τ_{wcell} is the delay associated with the cell state change given the applied BL differential and WL voltage. Simulations will be used to quantify τ_{wcell} in this study.

4.5.3 Margin/delay analysis

A margin sensitivity factor and performance factor will be employed to derive a final effectiveness factor and a graphical (margin/delay) space analysis will also be used [56]. To illustrate the margin/delay approach, Fig. 4.3 shows a schematic diagram depicting the desired functional window, delineated by the margin and delay requirements of the memory. As the VDD is reduced to Vwc, the read/write margins and corresponding performance degrade. Use of assist methods generally improves margin and in most cases

delay to some extent. Plotting the margin versus delay of a memory with varying amounts of assist bias will illuminate the most effective assist methods for a given technology and set of functional requirements. This graphical approach provides additional insight into the net functional impact of a given assist method and allows us to readily understand potential advantages and trade offs of a given assist approach.

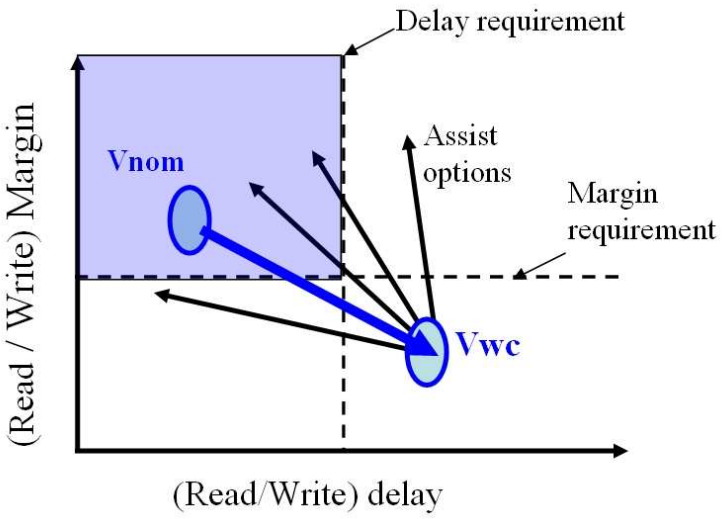


Figure 4.3: Schematic diagram of read/write margin vs read/write delay and desired functional window based on margin limited yield and performance requirements for application.

4.6 Results

Four read assist and four write assist methods were examined to provide a set of test cases for the assist evaluation methodology. A schematic representation of the specific assist methods explored is given in Fig. 4.1 and 4.2 for read assist and write assist respectively. Three of the read assist methods chosen for this evaluation were of type 2 category and one (WL droop) was type 1. Two of the write assist methods chosen for this evaluation

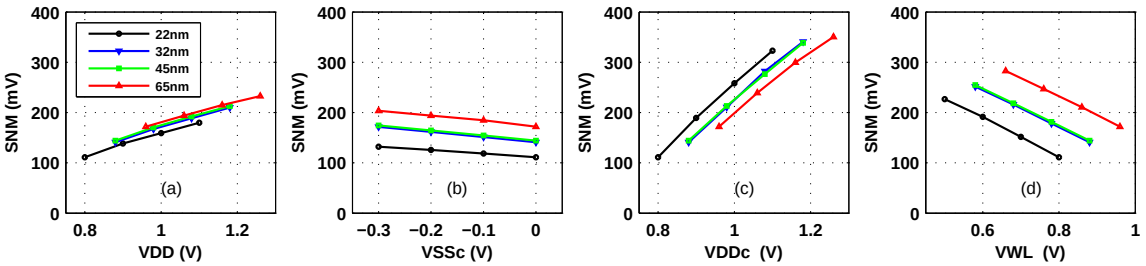


Figure 4.4: Read static noise margin as function of (a) raised array global VDD, (b) Negative VSS at the cell, (c) VDD boost at the cell (VDDc) and (d) WL droop.

were from type 1, and the remaining two were type 2. The four read assist methods shown are listed in Table 4.1 rows 1-4. The four write assist methods discussed in this work are given in Table 4.1 rows 2-5. Those assist methods that are inherently dynamic (influencing the duration of the noise source) must be evaluated using dynamic noise margin methods. These include reduced BL capacitance and read modify write or write back.

4.6.1 Simulation results - margin

To quantify the margin sensitivities in this study, static metrics will be used to emulate the functional environment using the custom low power (LP) PTM bulk technologies [12]. For read assist, SNM based on the butterfly curve analysis is used [74]. For write assist, the ramped WL based metric will be employed [25], defined as the (VWLmax - VWLflip) to assess the margin due to its improved correlation to dynamic write margin [85]. A yield analysis will be used to establish a quantitative relationship for the required margins.

Fig. 4.4(a-d) plots the SNM as a function of the assist bias for the four read assist techniques defined in Fig. 4.1(a-d). The four technology nodes are represented in each of the four plots. Fig. 4.4(c) for example shows the change in SNM with increased array VDD (VDDc) as described schematically in Fig. 4.1(c). There is a negative slope for methods (b)

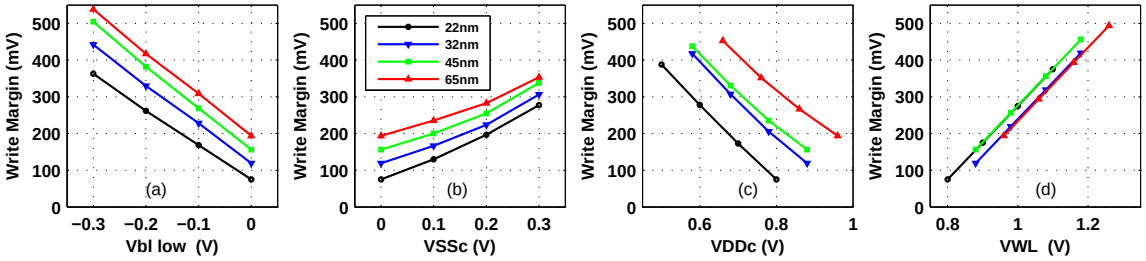


Figure 4.5: Write margin as function of (a) negative BL, (b) raised VSS at the cell (VSSc), (c) VDD droop at the cell (VDDc) and (d) WL boost.

and (d) corresponding with the fact that these methods utilize a reduction in the terminal voltage. While all four methods produced some degree of improvement in the SNM, and the response or sensitivity is similar across the technology nodes, the sensitivity was most significant for VDDc, Fig. 4.1(c) and Fig. 4.4(c). The initial voltage is either 0V or varies consistently with the Vwc for each technology.

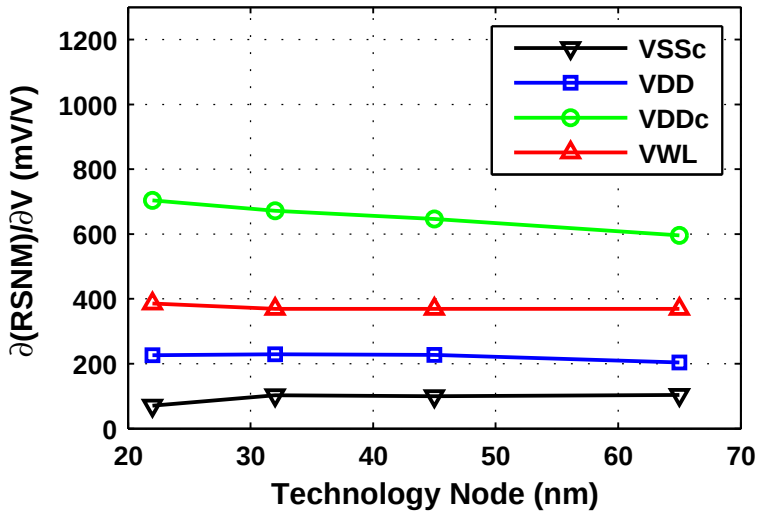
The simulation results for the write assist methods are shown in Fig. 4.5 (a-d) corresponding with the conditions defined in Fig. 4.2 (a-d). For the write assist methods in this analysis, the VSSc response, Fig. 4.5(b), was the least linear and showed the least sensitivity. Although there is some degree of non-linearity in the response characteristics of write margin and static noise margin, most exhibit a sufficient degree of linearity across the 300mV range to allow us to characterize the responses using a first order linear model to allow a high level comparison. SNM sensitivities shown in Fig. 4.4 (a-d) are summarized for each of the technology nodes in Fig. 4.6(a). As a means of improving the SNM, raised cell voltage (VDDc) is the method that emerges as exhibiting the greatest sensitivity across the LP technology nodes. The trends also suggest that there is some increase in sensitivity as scaling continues.

Following a similar approach, the functional sensitivities were also characterized across

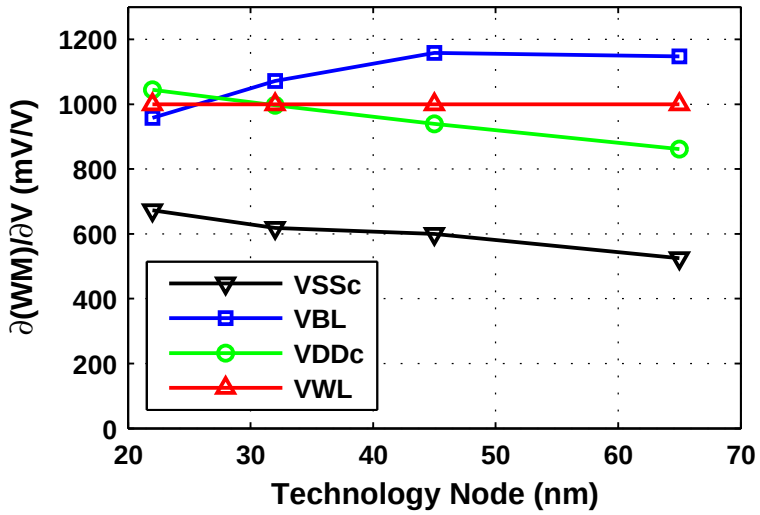
the technology nodes for write margin sensitivity, Fig. 4.6(b). In this case, three of the methods exhibit similar sensitivities to the applied bias. Raised array VSS (VSSc) showed less degree of linearity and had a weaker response. The unique and completely linear response of the WL boost was due to the fact that the write margin metric used in this investigation was defined as the difference between the final word line voltage and the voltage of the word line required to write the cell.

4.6.2 Simulation results - performance

The relationship between read current and read SNM is of particular concern with scaled technologies as the read currents are generally decreasing with successive generation. The read assist methods have an important and significant impact on the cell read current. The influence of the read assist methods on the read current for the 45nm node is shown in Fig. 4.7(a) with the initial value representing no assist technique at the low voltage corner (Vwc). Fig. 4.7(b) further plots the spread of read current vs SNM at 300mV assist bias. Although only the 45nm technology data is shown, the other three technology nodes responded in a similar way. Increased array voltage (VDDc) has only a small positive impact on the read current, while reduced word line voltage significantly degraded the read current. Decreasing the VSSc terminal below GND resulted in the strongest improvement in read current, exceeding that of conventional VDD increase. This results from both increased VGS and reduced threshold voltage in the SRAM cell pull down (PD) NMOS device due to the body effect. The read performance impact of the read assist techniques can be estimated for each technique with (2). Based on the simple relationship provided in (3), the performance limitations associated with the WL droop can quickly become prohibitive.



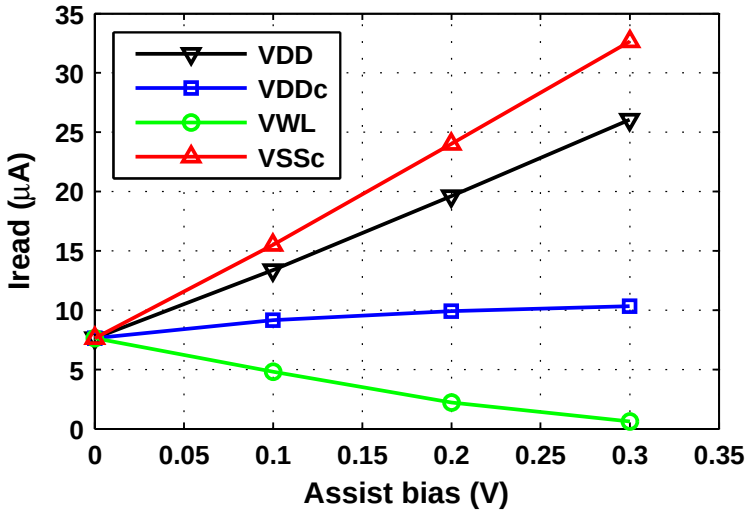
(a) SNM sensitivity by node



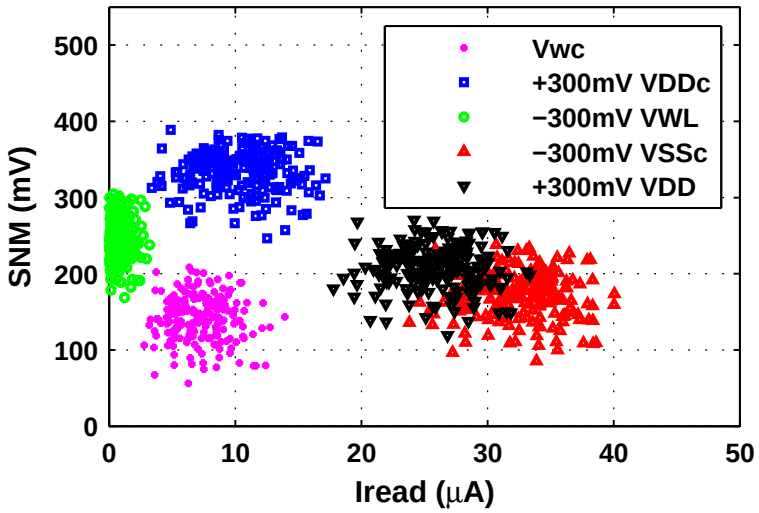
(b) WM sensitivity by node

Figure 4.6: The margin sensitivities across LP technologies for the four read assist methods (a) and four write assist methods (b) investigated.

The delay impact associated with the cell write time (τ_{wcell}) is shown in Fig. 4.8(a-d) for the four write assist methods evaluated. While all four methods improved the write time, WL boost and negative BL voltage bias schemes showed a more significant improvement



(a) 45nm LP read current vs assist bias



(b) 45nm LP SNM vs read current(Monte Carlo data)

Figure 4.7: The impact of read assist bias conditions on the bit cell read current (a) and SNM versus Iread for Vwc and 300mV of assist bias(b). Data shown is for the 45nm technology node.

in delay. Increasing the cell VSS and reducing the cell or array VDD had less impact. The delay response for cell write time was similar with scaling although the 22nm node showed

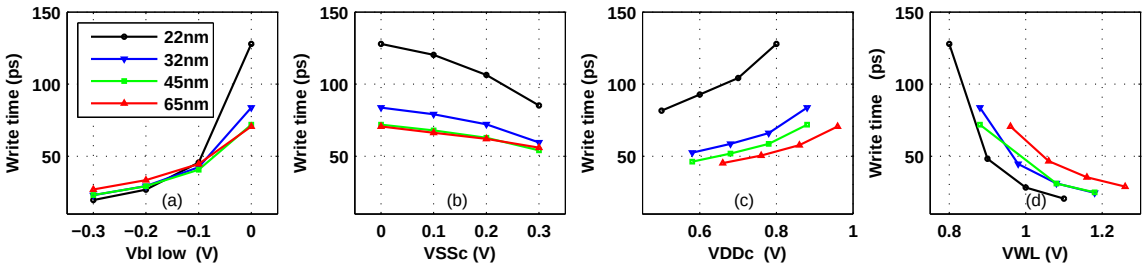


Figure 4.8: Effect of write assist techniques on cell component of write time (a) negative BL voltage, (b) raised cell Vss, (c) reduced VDD as the cell and (d) boosted WL voltage.

a stronger initial response to the applied bias conditions. For the negative BL and boosted WL cases, the 22nm delay response is most dramatically influenced by the 0.3V applied assist bias.

4.6.3 Impact of assist methods on variation

Until now, only the impact of the voltage deviations of the assist methods on the mean values of SNM and WM at a given bias condition have been discussed. However, to determine the functional yield expectation for a given array size at the worst case voltage, the local and global variation must be taken into account. Without the variation component, the required margin improvement will be unknown. For the small scaled SRAM devices, the local variation associated with random dopant fluctuations (RDF) dominates the variation sources. Although technology improvements offered by high-κ and metal gate, may provide significant improvement due to the higher gate capacitance, continued scaling will quickly consume these gains.

To address the impact of the assist methods on the variation in both SNM and WM Monte Carlo simulations were run for each method explored in this chapter. Fig. 4.9 plots the sigma for the WM distribution (a) and SNM (b) as a function of the assist voltage bias

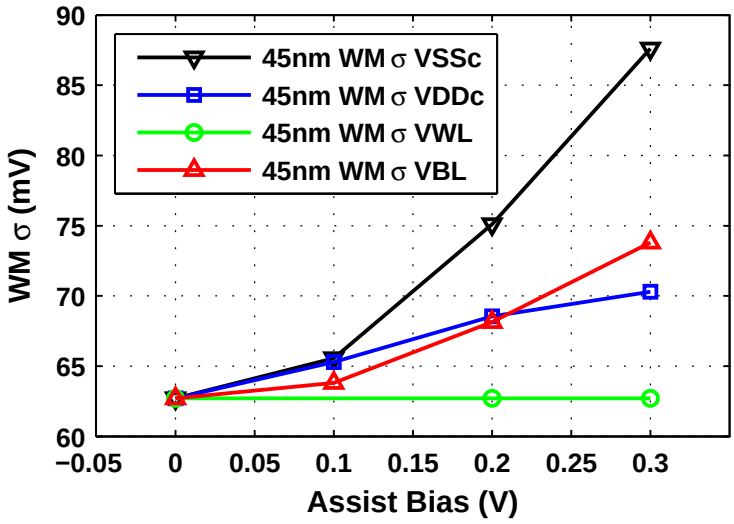
for the 45nm node. A minimum of 200 Monte Carlo runs were performed for each bias condition. Several observations emerge from this analysis. First, the assist method and bias both impact the standard deviation of the distribution. This is accounted for in assessing the overall contribution of the assist method which is discussed in the next section.

An additional source of variation in assist response can be caused by voltage variations on the assist modulated terminal(s). This variation will strongly depend on the specific design and assist implementation scheme used. The sensitivity metric, discussed in section 4.5.1, provides a means of assessing the overall impact of this variation source by relating changes in terminal voltage to margin.

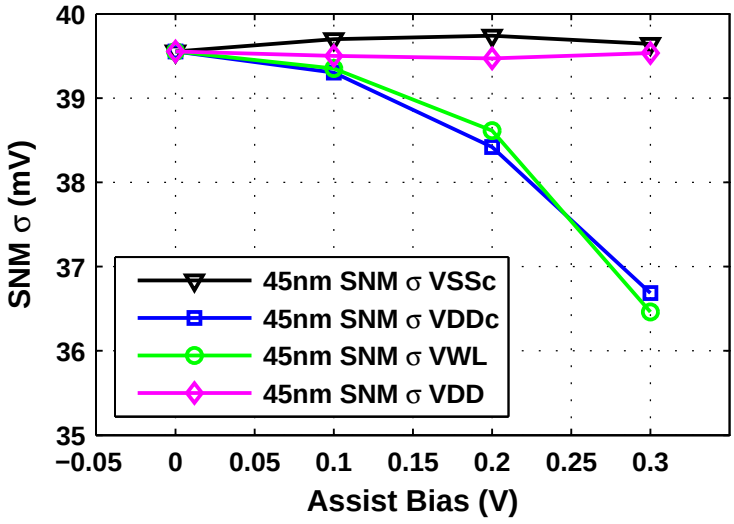
4.6.4 Yield Quantification

To identify the functional window requirement as depicted in Fig. 4.3, it is necessary to be able to convert the simulated margin information into yield. Soft fails are voltage, temperature, and timing dependent fails resulting from one of the following four modes: (1) failure to write, (2) failure to read (insufficient signal developed on the BL to set the sense amp), (3) stability upset, and (4) data retention. These four failure modes are not attributable to defects but are instead associated with a distribution tail stemming from variation sources. Although read fails and data retention fails are not addressed directly, assist method choices can clearly impact these mechanisms. The assist methods are directed at mechanisms 1 and 3. To address the write and stability related yields quantitatively, the following approach will be used.

SNM0/WM0 denote the read/write margin for data $'0'$ and SNM1/WM1 denote the margin for data $'1'$. The definition of SNM/WM would be the minimum value for $'0'$ and

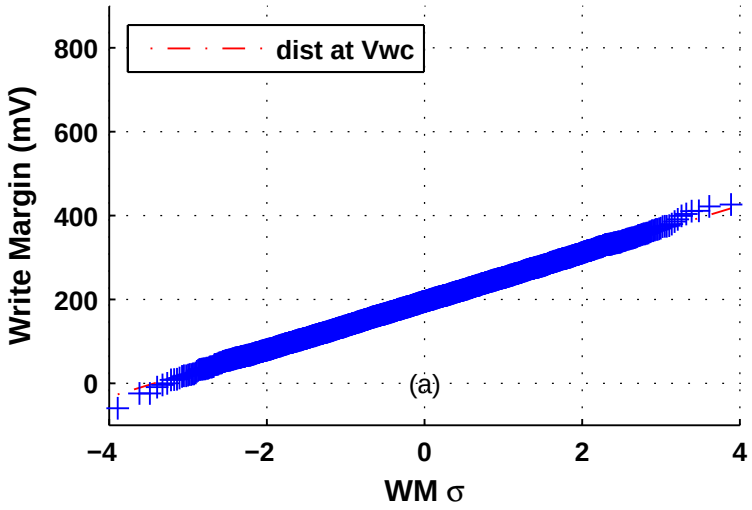


(a) 45nm LP WM σ vs assist bias

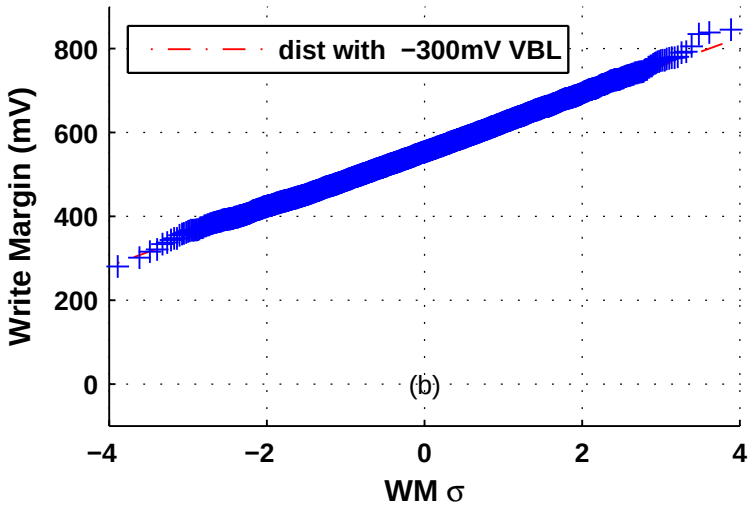


(b) 45nm LP SNM σ vs assist bias

Figure 4.9: Impact of assist method applied bias on the sigma of the resulting 45nm LP technology distribution for write assist (a) and read assist (b).



(a) WM(0) distribution at Vwc for 45nm LP



(b) WM(0) distribution at Vwc with -300mV VBL bias for 45nm LP

Figure 4.10: 10,000 Monte Carlo cases showing WM(0) standard normal distribution for 45nm LP technology at Vwc with no assist bias (a) and with 300mV negative BL bias (b).

$'1'$. An important observation is that the distribution of SNM0 or SNM1 can be represented by a standard normal distribution under normally distributed parameter variation. This same observation is true for WM0 or WM1. For the cases examined, the distributions remain normally distributed with assist bias, though the mean and the standard deviation may change. An additional set of Monte Carlo simulations (1,000 to 10,000 cases) were run on selected assist bias conditions for distribution verification purposes. Fig. 4.10 shows the results of 10,000 cases for WM0 at Vwc (a) and with 300mV negative BL bias (b) for the 45nm LP technology. The linearity of the quantile plots confirms that the WM distribution remains normal even with the assist feature engaged. The failure probability (Pf) for the right or left node (probability of SNM0¡0 or SNM1¡0 for example) is given as:

$$Pf = \frac{1}{2} erfc\left(\frac{\eta_\sigma}{\sqrt{2}}\right) \tag{4.4}$$

where η_σ is defined as the number of random variable standard deviations from the mean based on the standard normal distribution. For large arrays with relatively few fails, the Poisson distribution will be used to estimate the soft fail limited yield. (λ), defined as the number of bits (N) times the fail probability (Pf), can then be computed including both states of the latch:

$$\lambda = N \cdot (Pf_{(0)} + Pf_{(1)}) \tag{4.5}$$

With the assumption that the RDF induced variations are random and non-clustered, the soft fail yield (without redundancy) for a given mechanism can be expressed as:

$$Yield = exp(-\lambda) \tag{4.6}$$

To obtain a 10M-b SRAM with a SNM-limited yield of 99% would require a η_σ value of 6.12σ. In other words, to achieve this yield target, SNM0wc must be larger than the

minimum noise margin threshold (in this case 0) for 99 of 100 10M-b arrays. The limited yield for WM is computed with this same approach, to obtain a 99% WM-limited yield, which would result in an over all soft fail limited yield of 98% considering both WM and SNM. For our 45nm LP technology, Fig. 4.11 shows that this is achieved with 180 mV for either word line boost or negative BL bias (a) and 100mV assist bias for the most effective read assist technique (VDDc boost) (b).

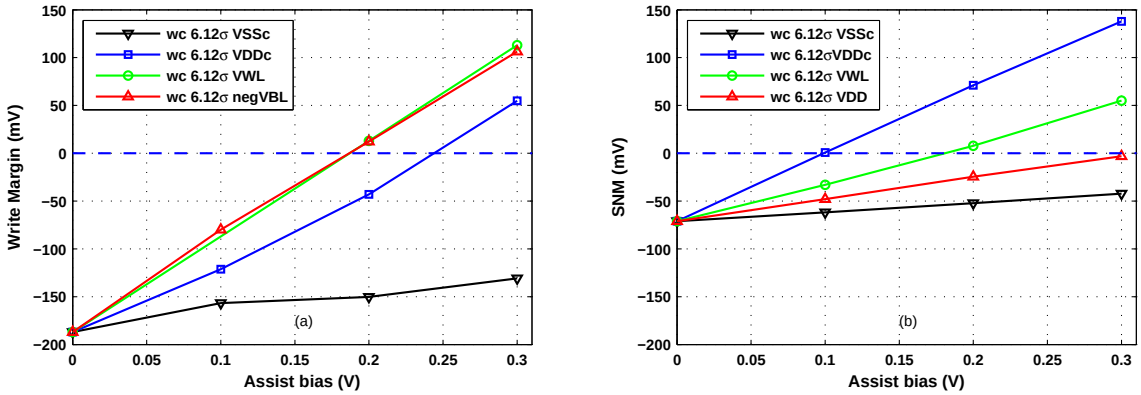


Figure 4.11: The 6.12σ worst case (wc) write margin (a) and SNM (b) as a function of assist bias for the 45nm LP technology.

4.7 Discussion

The elements of both margin and delay referenced to Vwc have been outlined. A means of transforming the write and read margins into a soft fail limited yield value has been provided. This approach has been applied and demonstrated using the LP PTM platform of bulk technologies from 65nm to 22nm.

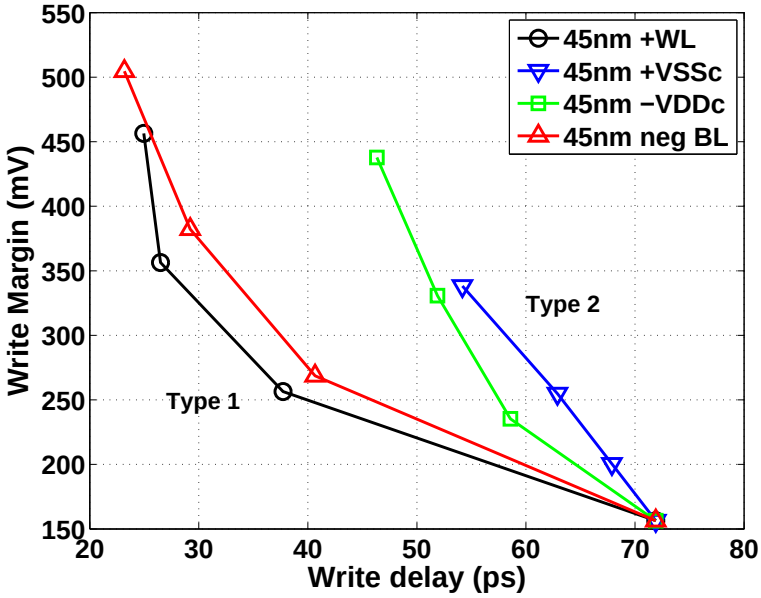
4.7.1 Assessing Functional Effectiveness

The functional read/write margin sensitivity was evaluated over a 300mV window to minimize non-linearity in the response and to ensure the bias conditions would not exceed the technology reliability limits. Because our reference (Vwc) condition was more than 200mV below nominal VDD in all cases, the reliability requirement was preserved. Even for the 22nm node where the Vwc was taken to be 0.8V, the max voltage would be only 10% greater than nominal VDD, which is consistent with common technology specifications.

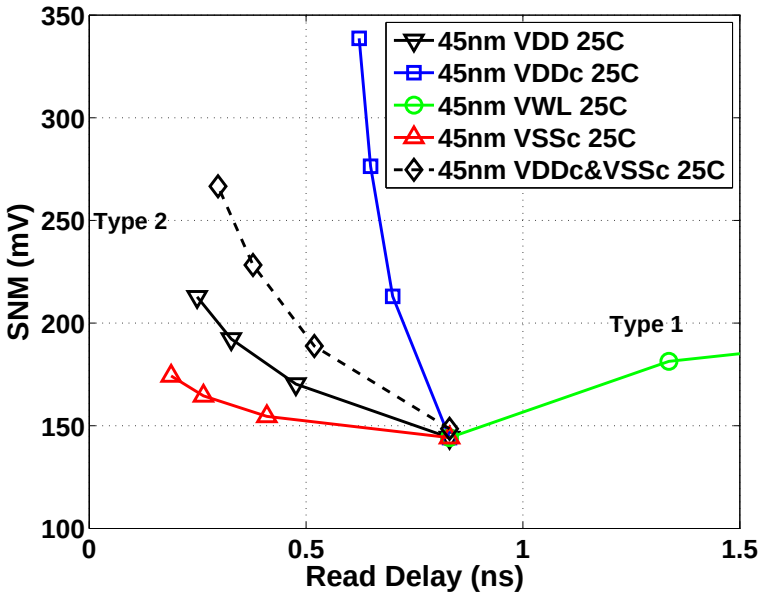
The sensitivity response for the assist methods studied is often influenced by more than one mechanism and can be understood when the device physics are taken into account. For example, the superior result associated with raised array voltage (VDDc) for read assist can be attributed to the fact that several mechanisms influence the result. The body effect causes the cell PFET to become stronger because of the modulated VSB for the PFET and the VGS is increased for the devices in the latch which are on.

4.7.2 Margin/delay space method

An example of the margin/delay plot introduced earlier is shown in Fig. 4.12 (a) showing the write margin versus write delay for each of the four assist methods evaluated. The different assist methods portray varying trajectories in the margin/delay space, and the type 1 methods are shown to increase margin while decreasing delay most effectively. Fig. 4.12 (b) shows the assist trajectories in margin/delay space for the read assist methods evaluated. A combined VDDc and VSSc assist method is shown in Fig. 4.12 (b) which demonstrates how the assist techniques can be combined as required to optimize both delay and margin. This figure also points out that some assist methods, such as WL droop, may improve the



(a) Write margin vs write delay



(b) Read margin vs read delay

Figure 4.12: Margin vs delay plots showing write (a) and read (b) for the 45nm LP technology when assist bias is swept from 0 to 300mV.

Table 4.2: Practical considerations for viable assist combinations

Read Assist Method	**Write Assist Method**	**cell compatible**	**low yield complexity**
Raise VDD	Raise VDD	yes	yes
-VSSc	+WL	yes	no
-VSSc	-BL	yes	yes
+VDDc	+WL	yes[a]	no
+VDDc	-BL	yes[a]	yes

[a] VDDc boost required for all columns on asserted WL

margin while simultaneously degrading the performance. Using this analysis approach, the methods categorized as type 2 were more effective for read assist.

The effect of variation was examined in some detail, and it was found that both assist method and bias had a non-negligible impact on the resulting WM and SNM distributions. For those cases where the assist method influenced the distribution, it was necessary to account for this in determining the effectiveness of a given method on the yield. While the SNM and WM distributions are intrinsically non-Gaussian for reasons previously discussed, relying on the distributions which are normally distributed, the distribution tail can be computed. By this method, a required assist bias for a given array size and soft fail yield requirement for both WM and SNM can be established.

4.7.3 Practical considerations

To assess the complexity of implementation for specific assist methods, yield implications associated with the specific assist method should be considered. For example, of the four write assist methods investigated, three (WL or VSSc boost, and VDDc droop) require a higher, yield related complexity. This is because WL boost increases the potential for stability upset in the cells along the asserted word line on the non-selected columns, and reduced voltage at the cell by VSSc boost or VDDc droop introduces data retention concerns. The trade off in the stability (SNM) impact of the half-selected bits during a write assist is shown in Fig. 4.13 for both negative BL and WL boost assists for 45nm. Although the negative BL method partially avoids these yield implications, the added level shift circuit complexity of generating the negative voltage must be considered.

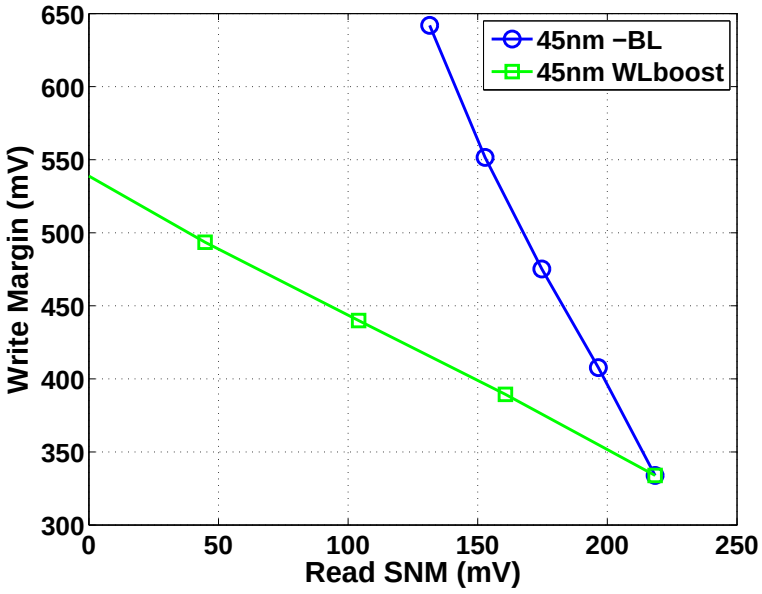


Figure 4.13: Impact of write assist on stability of the half-selected bits on the asserted word line shown for 45nm LP. As word-line-boost or negative-bit-line assist increases the write margin, the SNM is reduced for those bits on the word line subjected to a dummy read condition.

To address cell layout compatibility with a given assist method, it is noted that the 6T cell is typically provided by the foundry and therefore constrains the memory array designer to seek assist methods that best comply with the given layout. For example, the predominant industry 6T cell design style makes use of a VDD bus on metal level 2 (M2) level running parallel with the M2 bit lines. Although this layout style has advantages for density and performance reasons, the implementation of locally raising VDDc along the word line requires that all columns on the selected WL be boosted. Although pulsing the VSSc may be more consistent with this style cell layout (the metal 3 (M3) VSSc bus which runs parallel with the M3 word local line), this technique exhibited less margin sensitivity. It should also be pointed out that assist compatibility with dual port SRAM is of emerging importance, and some methods such as drooped VDDc for write assist are fundamentally incompatible. For those applications requiring both 1 and 2 port SRAM, the cost effectiveness for an approach such as the negative BL may become more compelling.

For those methods deemed most effective based on functional sensitivity and performance, the cell compatibility and yield complexity are considered together. Along with raised global VDD, four additional combinations of assist methods would need to be considered. Considering the predominant industry cell layout style, the comparison may then be summarized in Table 4.2. For the LP bulk technologies considered in this study, both read and write assist would be required to achieve high yield for large SRAM arrays beyond 65nm. Combining both the functional effectiveness requirement with the requirement that the cell layout must be compatible with the predominantly used industry bit cell, results in five pairs of options. By introducing the additional constraint that the yield complexity be low, the viable assist combinations reduce to three. For a final selection between the

remaining combinations of assist methods, absolute margin and performance deltas should be considered along with factors such as power and area overhead. An assessment of area overhead is dependent on the specific implementation scheme and therefore beyond the scope of this chapter, however, an area overhead of less than 4% would be expected for a competitive implementation [70] [66] [29] [64] [18].

4.7.4 Power

Power is a critical criteria for the ultimate selection of an assist method, however, power is dependent on both the assist method and implementation scheme. This is demonstrated by examining the essential components of SRAM array power. Both read and write operations are first described without assist and then for a specific read assist operation to illustrate this point.

The dominant components of power for a single read operation, without a circuit bias assist is given by:

$$P_{read} = P_{WL} + P_{\Delta BL} \tag{4.7}$$

where the components of read power are consistent with those in equation (4.2). The P_{WL} describes the power associated with the WL pulse and $P_{\Delta BL}$ refers to the power associated with the change in voltage on the BL's along the asserted WL. This read power may be expressed more fully as:

$$P_{read} = f(N_{BL}C_{WLc}V_{dd}^2 + N_{WL}C_{BLc}\Delta V_{BL}V_{dd}) \tag{4.8}$$

where f is the frequency, C_{BLc} and C_{WLc} are the bit line and word line capacitance per cell, N_{WL} and N_{BL} are the total number of word lines and bit lines in the array block of interest. The voltage differential required to set the sense amplifier is (ΔV_{BL}). The primary considerations for write power may be expressed as:

$$P_{write} = P_{BL\rightarrow 0} + P_{cell} + P_{WL} + P_{\Delta BL} \tag{4.9}$$

where the first three components of write power are consistent with those given in equation (4.3). Although not a contributor to write delay, $P_{\Delta BL}$ is a non-negligible component of the write power. The $P_{\Delta BL}$ term accounts for the power associated with the voltage change on the BL's along the asserted WL for the half-selected cells, i.e., those subjected to a dummy read operation. The $P_{BL\rightarrow 0}$ is the power associated with the BL discharge to ground for the write operation, P_{cell} is the power associated with writing the column selected cells on the word line, and P_{WL} describes the power due to the write WL pulse. The write power may be expressed more fully as:

$$\begin{aligned} P_{write} = f(&N_{SBL}N_{WL}C_{BLc}V_{dd}^2 + N_{SBL}C_{cell}V_{dd}^2 \\ &+N_{WL}N_{BL}C_{WLc}V_{dd}^2 + (N_{BL} - N_{SBL})\Delta V_{BL}V_{dd}) \end{aligned} \tag{4.10}$$

with N_{SBL} used to refer to the number of bit lines that are selected for the write operation.

A first order assessment in the change in power associated with a given assist method can be derived from these equations. For example, the change in power associated with the WL droop read assist can be expressed as:

$$\Delta P_{read} = f(N_{BL}C_{WLc}(V_{\Delta WL}^2 - 2V_{\Delta WL}V_{dd})) + P_{assist} \tag{4.11}$$

where $V_{\Delta WL}$ is the voltage reduction on the WL, P_{assist} is the power expended by the specific assist scheme chosen. The power associated with achieving the dynamic voltage reduction in the WL (P_{assist}), would also need to be included in the analysis. For example, the use of a replica or set of replica pass gate devices [66], which lower the WL voltage but also provide a DC path to ground during the WL pulse, would constitute a non-negligible P_{assist} when assessing the overall power impact. A similar analysis can be used for each assist method and implementation scheme. It is also clear from this analysis that the power will be dependent on specific array configuration factors, e.g., N_{BL}, N_{WL}, and N_{SBL}. In addition to the specific assist implementation scheme and array configuration, the cell and array layout configuration is also an important factor. For example, it would follow from this analysis method that the power impact of dynamically modulating the array supply bus for VDDc assist, with the conventional 6T layout and the array layout configuration discussed in section 6.3, may easily be large compared to other dynamic schemes.

A first principles analysis of relevant power components for both read and write without assist bias schemes was shown. Using this analysis it is also shown that determining the power for a given assist method requires specific details of the assist scheme and layout configuration. Because of the significant differences in margin sensitivity and performance across the assist methods, it is recommended that assessing the implementation costs and power be evaluated after determining the methods which are shown to satisfy the product functional requirements.

4.8 Conclusions

As competitive forces and industry scaling continue to erode the 6T SRAM functional margins, the use of assist methods will increase. A review and categorization approach for examining potential bias based assist methods is provided. For the assist methods evaluated in this study using the LP bulk CMOS technologies, those methods categorized as predominantly type 1 are more effective for write assist and the predominantly type 2 category of assist methods are more effective for read assist. The assist methods exhibited some degree of consistency across the platform of LP technologies studied. This suggests that the design infrastructure and assist method implementation cost can be reduced with reuse across multiple generations. The margin/delay analysis was demonstrated as an objective means of evaluating the influence on the functional metrics by the assist methods. Based on a margin/delay analysis and practical considerations, the more viable assist methods for future investment were identified, however, for a final selection additional factors such as implementation cost and power will need to be included in the analysis.

Chapter 5

Limits of Bias Based Circuit Assist Methods in Nanoscale SRAM

5.1 Introduction

Reduced device dimensions and operating voltages that accompany technology scaling have led to increased design challenges with each successive technology node. Large scale 6T SRAM arrays beyond 65nm will increasingly rely on assist methods to overcome the functional limitations imposed by increased variation, reduced overdrive and the inherent read stability/write margin trade off. Factors such as reliability, leakage and data retention establish the boundary conditions for the maximum voltage bias permitted for a given circuit assist approach. These constraints set an upper limit on the potential yield improvement that can be obtained for a given assist method and limit the minimum operation voltage (Vmin). By application of this set of constraints, it is shown that the read assist limit contour (ALC) in the margin/delay space can provide insight into the ultimate limits

for the nanoscale CMOS 6T SRAM.

5.2 Background and Motivation

Increased device variability and reduced overdrive associated with lower operating voltages have reduced the functional yield margins in VLSI circuits. This is particularly true for the 6T SRAM, which continues to play a dominant role in future technology generations because of its combination of density, performance, and compatibility with logic processing. Because of the commercial success of the 6T SRAM, methods to address the failure mechanisms of large memory arrays will extend the life of the 6T SRAM in VLSI circuits. Fail types for SRAM arrays may be divided into two distinct categories:"hard fails", i.e., those attributable to defects, and "soft fails". Soft fails defined in this context are those voltage, temperature and timing dependent fails resulting from one of the following four modes: (1) failure to write, (2) failure to read (insufficient signal developed on the BL), (3) stability upset during a read or half-select condition, and (4) data retention failure. These four failure modes each first occur at the distribution tail stemming from global and local variation sources.

The use of bias based circuit assist methods has become increasingly common, primarily to address soft fail modes 1 and 3 and to preserve the 6T cell functionality as the variation continues to increase and both read and write margins decrease with scaling. Although numerous recent articles have discussed bias based assist for SRAM as reviewed in chapter 4, limitations exist for all of these techniques. This limit may be reliability, performance, leakage, energy, or other factors which ultimately bound the extent to which the assist method compensates for the reduced functional margins.

The objective of this chapter is to explore the boundaries of bias based assist methods to understand the impact on the minimum operation voltage (Vmin) and the effectiveness of the assist methods for future generations of 6T SRAM. Based on the relationship between performance and functional margin with the applied bias constraints, the assist limit contour (ALC) for read assist is defined across four technology generations. For write assist methods, besides the constraint from reliability, the read stability of half-selected cells limits the permissible assist bias.

By application of the constraint limitations the maximum assist margin values permissible can then be mapped. The maximum permissible assist bias based on the reliability constraints are defined for each technology. The reliability limit may be due to several factors such as time dependent dielectric breakdown, hot carrier, NBTI or a combination of the known mechanisms with sufficient voltage acceleration. The maximum assist bias offset $|V_{assist}|$ that may be applied for any given assist method based on the reliability (Vmax) constraint is expressed as:

$$|V_{assist}| = (V_{max} - V_{nom}) + V_{droop} \tag{5.1}$$

Vnom and Vmax refer to the nominal and maximum operation voltage as specified by the technology developers (Vnom values provided in Table 1-1 of chapter 1 for this work). Vdroop refers to the difference between Vnom and the instantaneous operation voltage. To illustrate this concept briefly, for a technology in which the Vnom/Vmax is 1.2V/1.32V respectively, if the array VDD is drooped from 1.2V to 1V, the maximum assist bias is 0.32V. Any bias exceeding 0.32V would exceed Vmax for the transistor, violating the reliability constraint. For the same reason, a maximum negative assist bias of 0.32V may be

Table 5.1: Summary of constraints for bias based assists

Assist	**Method**	**Bias Constraint**
	WL voltage ↓	Vmax
Read	Pass Gate Vt ↑	Vmax
Assist	BL voltage ↓	Vmax and Write 0
	Array VDD ↑	Vmax
	negative VSS ↓	Vmax, Vfwd
	PMOS $\|Vt\|$ ↓	Vmax
	WL voltage ↑	Vmax, RSNM (1/2 select)
Write	negative BL ↓	Vmax, Vfwd
Assist	array VDD ↓	Vmax, DR (with shared VDDc)
	array VSS ↑	Vmax, DR (with shared VSSc)
	PMOS $\|Vt\|$ ↓↑	Vmax, RSNM (1/2 select)

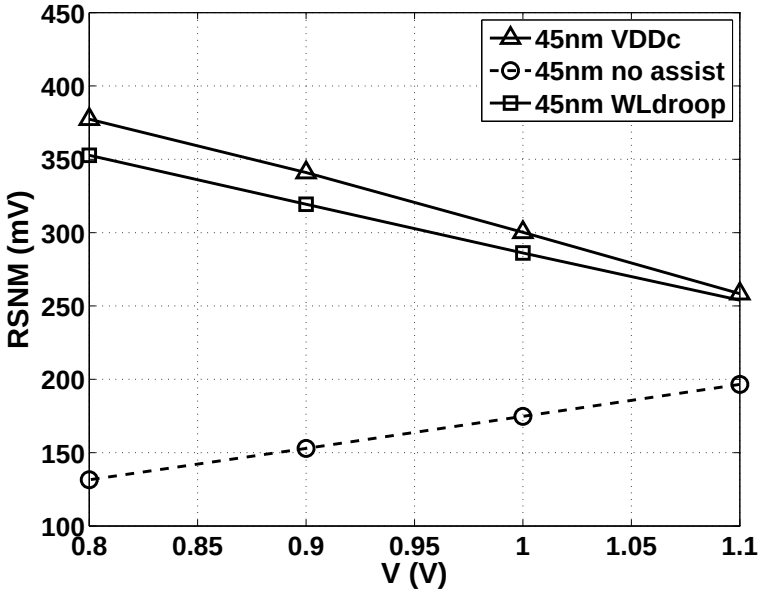
applied provided all VDD supply terminals associated with the array are maintained at 1V. Additional constraints may apply, but this single constraint provides a defined boundary that will be discussed further in section 5.3 of this chapter.

In addition to the technology defined Vmax constraint, other assist bias constraints for read assist bias include; forward bias diode turn-on (Vfwd) when VSSc is intentionally driven below ground, and cell upset by writing a zero when both bit lines are drooped sufficiently low (Write 0). For write assist, the constraints are again reliability (Vmax) as well as data retention (DR) for non-accessed cells sharing the intentionally modulated common supply, forward biased diode turn-on (Vfwd) when the 'write zero' bit line is driven below ground, and cell stability for the half-selected cells on the asserted word line.

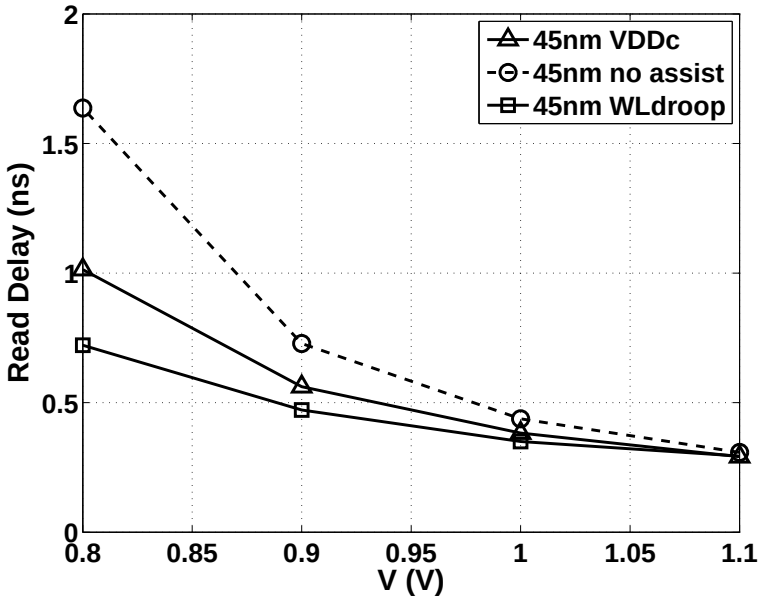
The primary bias constraints are summarized in Table 5.1 for the bias based assist methods evaluated in this chapter. Vmax is a valid constraint for all cases. This is less obvious for the two write assist options that involve collapsed supply across the latch. Vmax remains a constraint for the maximum write margin because it still limits the maximum WL voltage. For the purposes of this work, Vmax will be defined as 10% above the nominal operation voltage. Because Vmax is a limiting factor in all bias based assist methods, this fact may be exploited to explore the limits of the assist methods across the scaled technologies. This approach allows us to effectively define the upper envelope of assist bias conditions permissible for a given technology. By mapping the assist methods across the margin/delay space, the functional window may then be used to illuminate the practical voltage bias boundaries.

5.3 Results

To examine the maximum soft fail limited yield boundaries that can be achieved for a given assist method, the relationship with VDD is first described and then applied to the assist bias using the Vmax constraint. The read static noise margin as a function of VDD is shown in Fig. 5.1 for the 45nm LP PTM technology. The three cases shown are with array VDD (VDDc) boost, array VSS (VSSc) reduced, and with no assist. It is clear from Fig. 5.1(a) that the use of the maximum assist bias, consistent with relationship (5.1), can significantly improve the otherwise reduced static noise margin when the word line is asserted. The SNM improves beyond the nominal value when the VDDc assist method is invoked because the noise source is being reduced with VDD reduction, and the latch strength is increasing with the boosted VDDc.



(a) RSNM vs VDD.



(b) Read delay vs VDD.

Figure 5.1: Change in RSNM with reduced VDD (a) and effect of VDD on read delay (b) with the maximum allowable assist bias at each VDD. Data based on 45nm LP PTM.

Competing mechanisms produce a different result with negative VSSc. In this case, although the net latch strength is improved over the non-assist case, the noise source through the pass gate NFET is becoming stronger due to the body effect producing a reduction in pass gate VT on the side of the cell storing a zero. Additionally, the VT is reduced for the pull down NMOS device with drain storing a one. This results in an earlier turn of this pull down NMOS and further reduces the SNM. The read delay for the cell is improved due the body effect which strengthens both the pull down and pass gate series devices on the side of the latch storing a zero.

While the cell stability compensation is larger for VDDc assist, the improvement in performance (read delay) may not be sufficient depending on the functional window as discussed earlier. Boosting the VDD at the cell (VDDc) has a small impact on the read delay, consequently the read delay continues to degrade as VDD is reduced. The alternate read assist method (VSSc) shown in Fig. 5.1 improves SNM to some degree but more significantly improves the read performance. This is because the body effect associated with reduced VSSc causes both the pull down and pass gate NFET device VT to be reduced, boosting the read current.

The margin/delay relationship is applied for the assist methods with maximum assist bias. The effect of maximum assist bias on both SNM and delay based on the modulation of single and multiple terminals is shown in Fig. 5.2 for the 45nm LP-PTM technology. Each of the read assist bias conditions given in Table 5.1 except those involving well bias VT modulation were employed.

The margin/delay analysis reveals the limits of the bias based assist methods across the relevant design space. This boundary further defines a contour, as shown by the solid

continuous line (demonstrated using VDDc and VSSc assist bias following the Vmax constraint). The boundary referred to as the assist limit contour (ALC). It establishes the effective limit in SNM and corresponding relationship to read performance for a given technology and bit cell. This boundary or ALC mapped by the assist methods therefore provides a means of assessing the functional limits of the 6T SRAM.

For the cases studied, the read ALC as defined by the latch supply voltages were found to provide a reasonable approximation of the full multi-terminal Vmax read assist contour. Because drooping the WL provides a degree of freedom that is not limited by the Vmax constraint, those combinations of negative VSSc combined with WL voltage reduction where found to produce a slightly improved margin/delay response for the LP-PTM technologies.

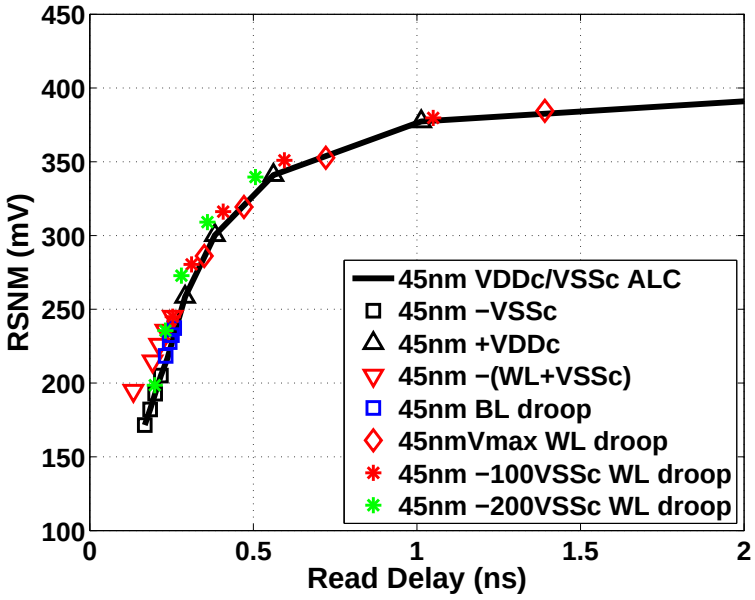


Figure 5.2: Multiple read assist options involving both single and multiple terminals with Vmax constraint preserved.

Because the primary goal of this work is to identify and delineate the bias based assist

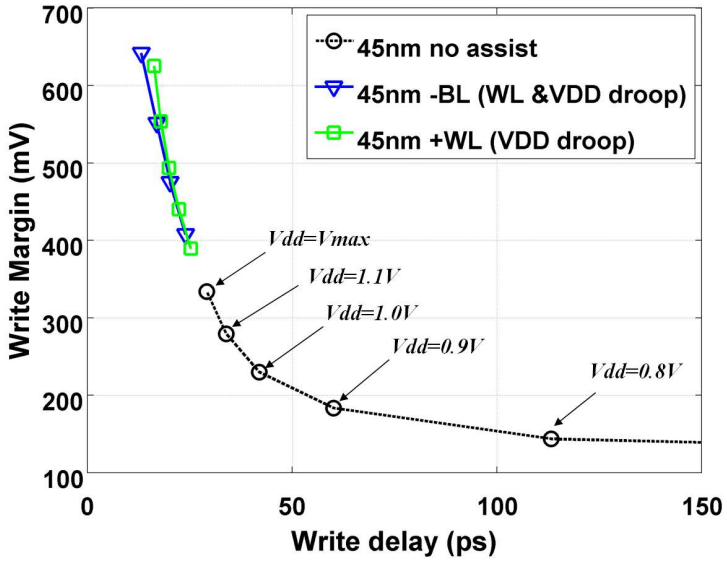


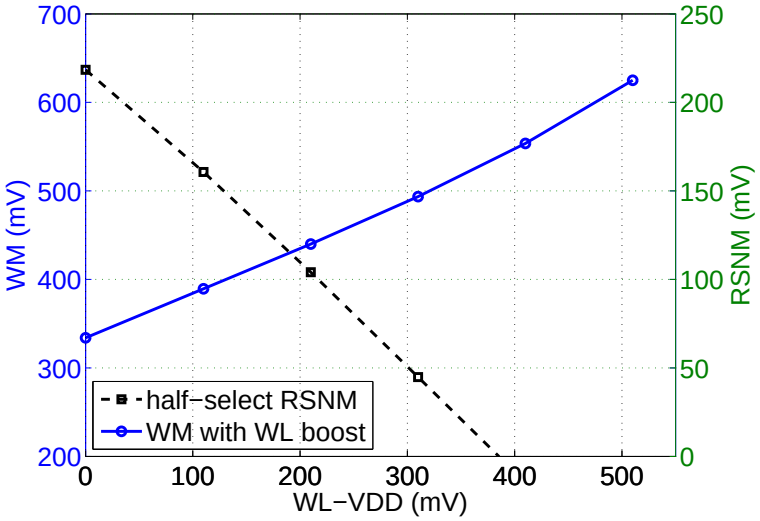
Figure 5.3: Write margin decreases as VDD is reduced when no assist is used. With assist at Vmax, the write margin is increased with reduced VDD.

limitations of the scaled 6T SRAM cell, the delay required in developing the bias conditions is not included in this analysis. A complete SRAM macro design would need to include the overhead delay associated with the specific implementation and circuit choice.

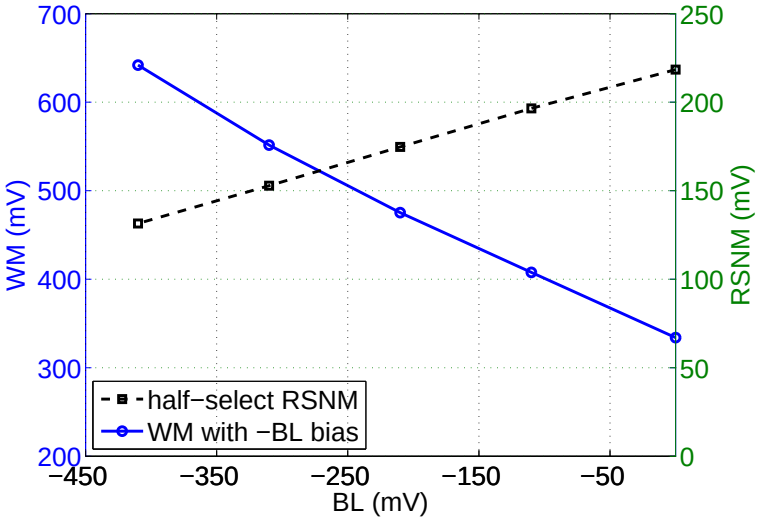
For write assist, the margin/delay analysis leads to a different result. In this case, there is no inherent trade off between write margin and write delay. The relationship is shown in Fig. 5.3 for two assist methods (negative BL and boosted WL) showing improved margin and delay as array VDD is drooped. For boosted WL assist, the array VDD, high bit line and NWELL voltages are reduced, while the WL line is boosted to Vmax limited by the reliability constraint between the WL voltage and the low (write zero) bit line at 0. For the negative BL case, as the VDD is reduced on the word line, high bit line, and array VDD, while the low bit line is drooped by the same amount to preserve the Vmax constraint. With the word line boosted to Vmax, the write margin continues to increase

with corresponding VDD reduction. Similarly, with the (write zero) bit line driven below ground by a value equivalent to the VDD reduction (preserving the Vmax constraint), the write margin continues to increase. In addition to Vmax, the maximum write assist bias may become limited by other constraints, such as the read margin for the half-selected bits, shown in Table 5.1.

For the WL-boost write assist, Fig. 5.4(a) shows that as the array supply voltage is reduced, boosting the WL while preserving Vmax reduces the stability (RSNM) of the half-select bits on the same WL. Therefore, the limiter for the WL boost quickly becomes the reduced SNM on the half-selected bits. For the negative BL assist, the BL bias does not directly impact the half-select bits. However, because the amount of bias between the BL voltage and the global VDD is limited to Vmax, a larger negative bias on the BL implies a lower global VDD. Thus the RSNM of the half-selected bits consequently decreases, permitting a larger negative BL bias, as shown in Fig. 5.4(b). By comparison, the degradation in RSNM for the half-selected bits using the negative BL assist, Fig. 5.4(b), results in less degradation for the half-selected RSNM. This is an advantage of the negative BL assist over the WL boost. However, as the array supply droops, the negative BL bias can eventually become limited by leakage to the substrate as the forward bias diode begins turning on. To overcome the problematic stability concern for the half-selected bits during a write, a read assist such as VDDc boost may be applied to the non-selected columns. Alternatively, the array architecture can be designed so that the half-select is avoided and all bits on the asserted WL are latched during a write operation.



(a) WL boost.



(b) Negative BL.

Figure 5.4: (a) The impact of WL boost on the WM of the selected bits and the stability (RSNM) of the half-selected bits. (b) The impact of negative BL on the WM of the selected bits and the stability (RSNM) of the half-selected bits.

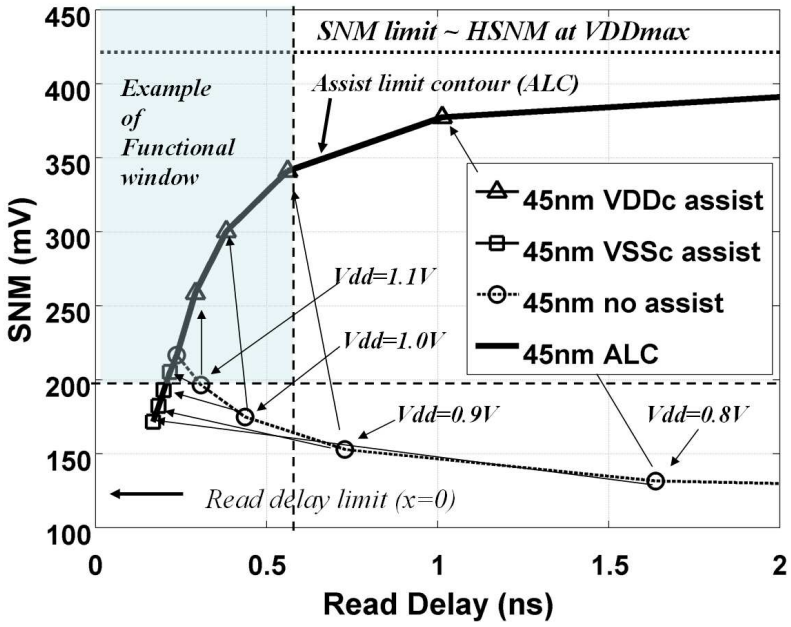
5.4 Discussion

The characteristic features of the margin/delay plot for read assist were examined. Fig. 5.5(a) reveals that the read assist limit contour (ALC) asymptotically approaches the hold SNM (HSNM) limit with increased delay. A simple model is used to describe the observed contour shape. With some simplification, as defined in Appendix C, the read delay can be approximated by the following relationship:

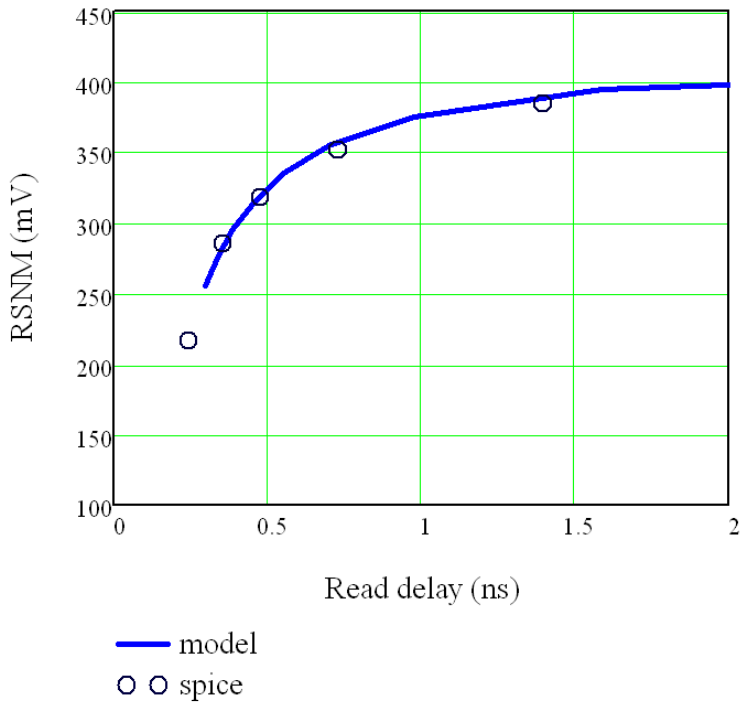
$$\tau_{read}(Vwl, Vddc, Vtn_0) = \frac{2C_{BL} \cdot L_{pg}\delta V_{BL}(\Psi_n\gamma_x - \Psi_n\gamma_n - Vtn_0\beta\gamma_x + Vddc\beta\gamma_x)}{W_{pg}\Psi_n\beta k_n\gamma_x(Vtn_0 - Vddc) \cdot (2Vtn_0 - 2Vwl + \Psi_n)} \tag{5.2}$$

Where Ψ_n represents the velocity saturation value, γ_n is the body coefficient, γ_x is a linear approximation factor of the body effect as Vsb increases. Vtn_0 is the NMOS threshold voltage with Vsb=0 and is the PD NMOS W/L divided by the PG NMOS W/L value. The RSNM can be described as linear relationship with bias using the empirically derived sensitivity value obtained in chapter 3. The resulting analytical solution is given in Fig 5.5(b).

This general relationship may be anticipated as the latch strength is increased relative to the noise source, the SNM upper limit will approach the HSNM with VDD=Vmax. For the case where the NWELL potential is tied to the array VDD (VDDc), the upper limit will be equal to the hold SNM. This can be more clearly seen from the butterfly curves. Fig. 5.6 plots the butterfly curves when the VDDc assist method is used with increased assist bias. The characteristic shape evolves with increased assist bias, becoming more similar to the hold SNM shape. Because the performance implications of achieving this limit are in most cases not practical, the more relevant portion of the ALC is across the intersection of the



(a) Simulated assist limit contour.



(b) Assist limit contour based on analytical model.

Figure 5.5: (a) The VDDc/VSSc defined read assist limit contour (ALC) as defined by the margin/delay space for 45nm LP PTM 6T SRAM. (b) Analytical ALC model derived using SNM sensitivity with read delay, as calculated by (5.2).

functional window as shown schematically as a shaded region in Fig. 5.5(a).

For a given set of technology bias constraints, a contour line defining the upper most noise margin at a given read delay for the technology may be derived. Fig. 5.7 plots the read ALC mapped across margin/delay space for the LP-PTM technologies from 65nm to 22nm node. Note that for each technology node, the ALC exhibits a similar shape. By applying the specific functional window as determined by the use conditions, array size, and yield requirements, one may follow this approach to assess the viability of the bias based assist methods based on the overlap of the ALC and functional window.

For designs requiring both read and write assist, the yield limiting condition, if not otherwise addressed, may then be the stability upset half-selected bits during a write operation. As shown in Fig. 5.4(a) and

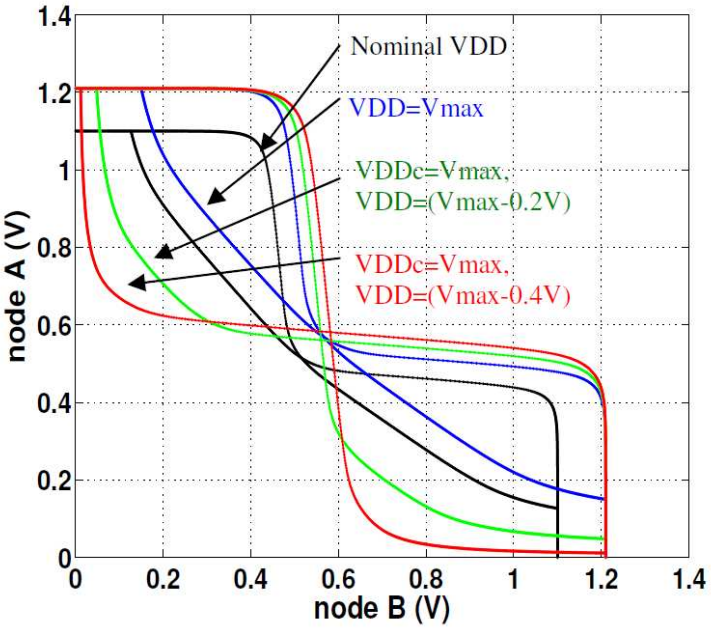


Figure 5.6: Simulated butterfly curves for nominal, Vmax and two VDDc assist cases from a 45nm LP commercial technology.

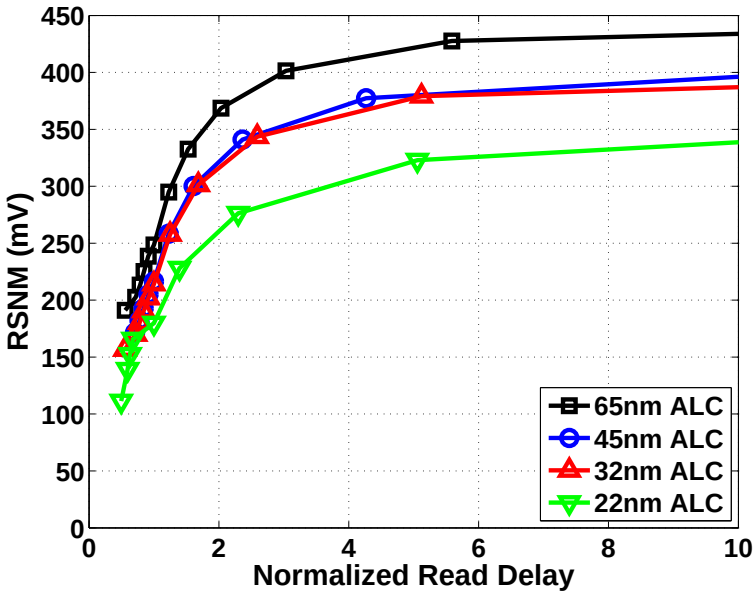


Figure 5.7: Simulated butterfly curves for nominal, Vmax and two VDDc assist cases from a 45nm LP commercial technology.

5.5 Conclusions

Continued scaling of the planar 6T SRAM will necessitate increased reliance on assist methods to overcome reduced functional yield margins. Because added assist features will incur costs in design complexity, area, and in most cases power, these factors must be balanced against the potential improvement in soft fail limited yield margin and performance. For bias based assist methods, bias constraints ultimately limit the margin improvements that can be obtained. The applied voltage bias associated with a given assist method must conform to the existing technology bias constraints.

For write assist, in addition to the Vmax constraint, other combined factors also limit the attainable margins. For read assist, by imposing the Vmax constraint a contour is observed in the margin delay space that reflects the relevant attainable limits of a given assist method. The intersection of the ALC with the functional window requirement provides

a means to establish bias based assist limitations for a given technology and bit cell. By accounting for these factors, the Vmax constrained read ALC is mapped across four technology generations to gain additional insight into the extent to which assist methods may continue to compensate for the reduced functional margins with continued scaling of the planar 6T SRAM.

Chapter 6

Summary and Conclusion

6.1 Summary of contributions

As scaling continues and both voltage and device dimensions are reduced, the functional window for SRAM is reduced, and the future of the 6T bit cell is less certain. In this dissertation, we address both sources of SRAM device variation and circuit methods of coping with variation as the technology interactions with circuit optimization are explored. The specific contributions of this work include:

Random and non-random mismatch considerations in the bit cell design environment

Several technology offerings have been proposed in the literature to address or offset the increased variation associated with random dopant fluctuations. These include high-κ gate dielectric materials and metal gate, ultra thin (UT) or fully depleted (FD) SOI technologies and non-planar solutions such as FINFET, MUGFET and surround gate technology solutions.

Circuits to enable statistical analysis and evaluation of the A_{Vt} for a metal gate FDSOI

technology were implemented and the hardware was analyzed to extract an $A_V t$ value for the PMOS devices of 2.4mV-μm to examine a leading technology solution for the random component of device variation. This result further supports the advantages of these emerging technology solutions as an enabling path to future scaled CMOS.

A detailed examination of device variation sources in the SRAM cell environment is presented. A description of how dopant fluctuations in nanoscale SRAM devices may be attributed to both random and non-random components. Three factors which play a role in the susceptibility to sources of non-random dopant variation are; 1) SRAM cell layout topology, 2) process scaling practices, and 3) pushed design rules used in dense SRAM bit cell designs.

Four specific sources of non-random dopant driven threshold mismatch that can arise in the SRAM device environment are; (1) implanted ion straggle in SiO_2, (2) polysilicon inter-diffusion driven counter-doping, (3) lateral ion straggle from the photo-resist and (4) photo-resist implant shadowing. This work is believed to be the first to highlight and address these mechanisms in the context of the aggressive bit cell design environment.

A new 6T planar bit cell topology for sub 20nm lithography

A re-examination of the fundamental layout options for the planar 6T SRAM bit cell, coupled with the increasing lithography constraints, lead to the exploration and proposal of a new family of cell topologies. The new bit cell topology offers three distinct advantages over the existing industry bit cell that is widely used today. It provides 1) reduced lithographic wiring complexity, 2) eliminates jogs in the active silicon for reduced contribution of geometric variation sources, and 3) offers shorter bit lines over the dominant industry bit cell used today, 4) potential for fully routed array with only 2 levels of metal. A provisional

patent titled “Improved Dense 6T SRAM Cell Layout Structure and Related Method” has been submitted on this new bit cell design topology [59].

The Margin/delay analysis metric

The primary focus of the circuit assist methods has been improved read or write margin with less attention given to the the implications for performance. In this work, margin sensitivity and margin/delay analysis tools are introduced for assessing the functional effectiveness of the bias based assist methods. A margin/delay analysis of bias based circuit assist methods is presented, highlighting the assist impact on the functional metrics, margin and performance.

A new method for concurrently optimizing the impact of circuit assist methods and biases is presented known as the margin/delay method. The concept of margin sensitivity is developed and discussed as a component of the margin/delay concept. The analysis spans four generations of low power technologies to show the trends and long term effectiveness of the circuit assist techniques in future low power bulk technologies. A publication titled “ Impact of circuit assist methods on margin and performance in 6T SRAM” was published in the Journal of Solid State Electronics [57].

Examining the limitations of bias based assist methods

Although circuit assist schemes provide improved yield margin for scaled SRAM, factors such as reliability, leakage and data retention establish the boundary conditions for the maximum voltage bias permitted for a given circuit assist approach. These constraints set an upper limit on the potential yield improvement that can be obtained for a given assist method and limit the minimum operation voltage (Vmin). By application of this set of constraints, it is shown that the read assist limit contour (ALC) in the margin/delay space

can provide insight into the ultimate limits for the nanoscale CMOS 6T SRAM. A paper titled “Limits of bias based assist methods in nanoscale 6T SRAM” was published in the proceedings from 11th International Symposium on Quality Electronic Design [56].

6.2 Extended work

Further investigations of cell layout topologies, process scaling, and pushed design rules on local variation within the SRAM bit cell will continue to be an important and valuable area for further research. Additionally, as circuit assist methods become more common, further research to address the trade-offs of specific implementations in power, performance and margin improvement are needed.

Extended work addressing sources of random variation in SRAM cell devices

1. In addition to RDF an additional source of random variation, which can be observed in small CMOS devices, is random telegraph signal (RTS) noise [24]. This noise source is characterized as a time dependent, low frequency variation that is of particular concern for narrow CMOS devices as used in the SRAM bit cell. Some work has been done by others to characterize the potential effects of this mechanism on Vmin at 90nm [3] and 45nm [82]. Additional work using the existing test setup and circuits characterize the effects for the FDSOI technology could provide more insight into the impacts of this source of random variation for future SRAM in FDSOI technologies.

2. The existing test methodologies and circuits can be further extended to stress and characterize the NBTI mechanism and characterize its impacts on the FDSOI technology.

3. Characterization of the A_{Vt} values for FDSOI or UTB SOI technologies will continue to be of great interest as the industry continues to seek the optimum technology and circuit design solution path for the next generation. As a planar solution, offering improved device variation and potentially improved layout density, this technology path is extremely promising. Additional characterization of the NMOS devices will be a valuable complement to this initial work which characterized 150nm FDSOI PMOS devices.

Extended work addressing sources of non-random variation in SRAM cell devices

1. In addition to the ideal layout structures (as employed on the MITLL FDSOI test chip), additional structures to enable the characterization of within-cell mismatch for a statistically significant number of SRAM devices (captured in the layout environment utilizing pushed design rules) would provide a natural extension of this initial investigation.

2. A statistical study of the within-cell variation as a function of alignment tolerance for each mechanism highlighted, coupled with the measured relationship to Vmin comparing two cell topologies would provide additional insight.

3. It is asserted by this work that the within cell mismatch for the entire (multi-lot) population will appear normally distributed about a mean mismatch value of zero. The non-random Vt mismatch will be apparent in examining the variation within a given lot (or appropriate alignment specific groupings) obtaining sufficient data to validate with statistical significance.

Extended work for circuit assist

Circuit assist methods are still only beginning to be optimized and adopted for high volume commercial SRAM applications. Although new methods and tools for characterizing and defining the optimum assist method have been introduced by this thesis, exciting work remains in this area of research.

1. The margin/delay analysis outlined in this thesis may stimulate further research in selecting the next generation assist methods. This work could be extended to generate a comparison and relationship of dynamic noise margin/delay with static noise margin/delay across several assist methods.

2. The versatility of the margin sensitivity metric may be exploited further in several ways. Additional work can be carried out towards exploring how the margin sensitivity metric can be used to provide insight and guidance for assessing the power impact of various assist methods.

3. Hardware measurements of SNM and WM across a range of voltage and temperature and technology platforms would allow correlation to simulations.

4. An observation that arose in studying the behavior of various assist methods was that the variation in SNM and WM could be modulated by specific types of assist methods. Specifically, hardware corroboration of this simulation result coupled with the development of a rigorous theoretical explanation for this simulation result would be a valuable extension of this effort and contribution to the field.

Extended work to explore the new 6T cell topology

By examining the layout implications of the industry-wide 6T SRAM cell, and characterizing the sources of non-random variation that can impact the dense SRAM devices, a

new bit cell topology was proposed that is expected to possess some advantages over the existing industry standard cell. Much work remains to fully develop this proposed layout topology.

1. Extending this work would involve incorporation of the type 5, 5e, or 5b layout topology at or below the 22nm node (using optimized pushed rules) and exploring alternative lithography options to continue to push the layout density of this topology.

2. Hardware based characterization, including yield and variation comparisons (following the work in item one above) would be a natural extension of this effort.

6.3 Conclusion and Outlook

SRAM has been and continues to be a technology driver, qualification vehicle, and competitive benchmark for logic and microprocessor technologies. Most recent estimates place the semiconductor industry revenue for 2010 to be on the order of $300Billion [83], with logic and microprocessors comprising 21 and 14.4% respectively [36]. Embedded SRAM comprises the bulk of the L1, L2, and L3 cache for today's microprocessors and is extensively used in ASICS and logic applications, where it is expected that as much as 50% of the chip die area may be comprised of SRAM. Extending the well developed planar 6T SRAM technology is therefore of enormous economic importance.

Circuit design complexity and challenges are increasing with each new technology generation. This thesis has focused on device variation (both random and systematic sources), characterizing the variation impact on circuits and developing solutions to address the im-

pact of variation in SRAM.

Specific solution paths explored in this work include both circuit and technology. The technology solutions were explored in two primary areas; 1) random variation, 2) systematic variation. To explore one potential solution path addressing random variation, test devices were designed and tested in a FDSOI technology. The improved device variation associated with the FDSOI or UT SOI, does show promise, however more statistical data will be needed.

Sources of non-random variation associated with scaling, topology and extensive use of pushed design rules in advanced SRAM were examined and characterized. A detailed evaluation of the bit cell topology options, and the implications on systematic variation was performed. Following this examination, a new category of layout topology was proposed which may provide and stimulate further investigation in this area. The new topology, while offering certain advantages over the industry standard 6T, does not achieve the same density when existing pushed layout rules are applied. Because of the dynamic and rapidly evolving technology options which are emerging, such as pitch doubling solutions and replacement gate process options, an additional topology category for further exploration may promote renewed interest in this area.

An objective, metric based methodology for examining and characterizing SRAM circuit assist methods is provided in this work. By examining the range of approaches and benefits of circuit assist methods, a method of categorizing the assist types was developed. A new margin/delay analysis method was developed to provide circuit designers with an objective means of better trading off the benefits of yield margin and performance impacts associated with the assist method.

A recognition that the bias based assist methods require voltage modulation of one or more terminals for the SRAM was the foundation for an exploration of the limits of this type of circuit assist. By a careful analysis of the limiting applied biases and the margin sensitivity associated with a given assist method, limits in the performance and margin gains can be established.

Despite the wide range of technical challenges outlined in the introductory section of this work, with the incorporation of both process technology and circuit innovations, the outlook remains optimistic for the next generation SRAM. Continued planar scaling beyond the 22/20nm node and perhaps as far as the 11/10nm node is anticipated.

Appendix A

Chip design

A.0.1 MITLL 150nm ULP FDSOI chip

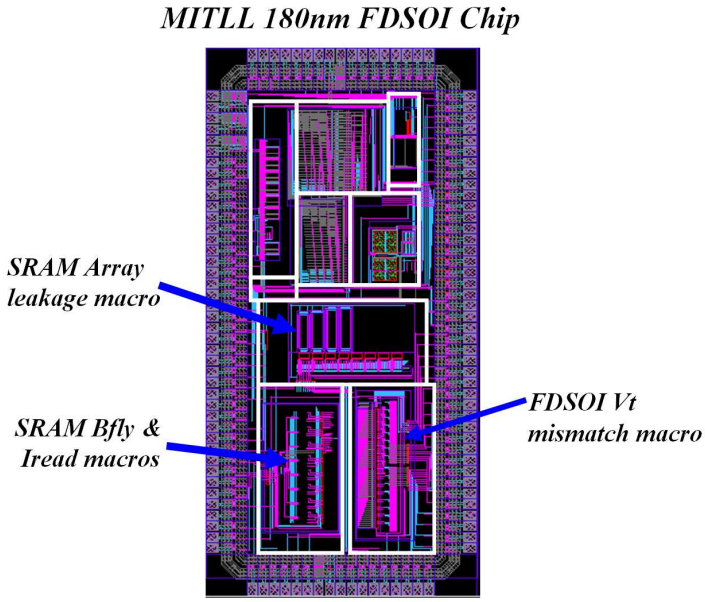


Figure A.1: MITLL 150nm fully depleted SOI technology chip design. Digital decoder design enables multi-device NMOS and PMOS device mismatch characterization, multi-array bit cell leakage during standby and butterfly curves and cell read currents from multiple SRAM bit cells.

A.0.2 MITLL 150nm FDSOI chip (die photo)

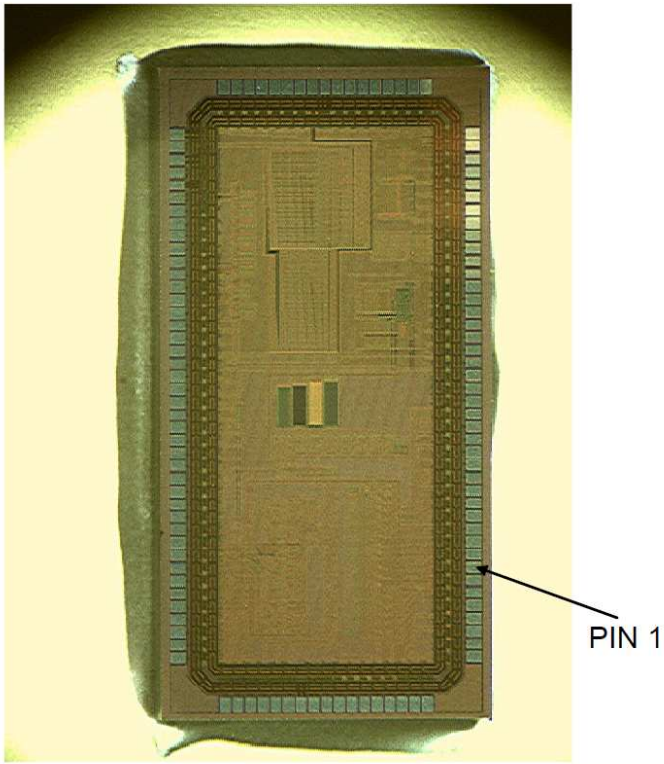
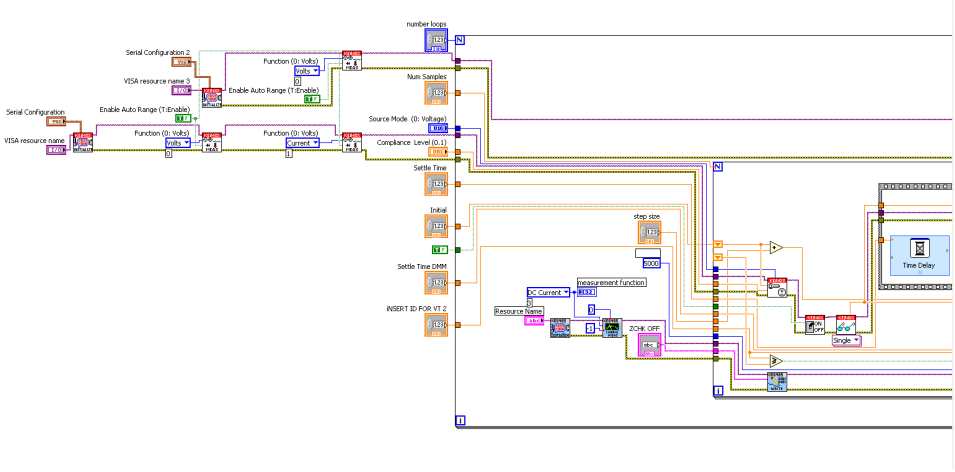


Figure A.2: MITLL 150nm fully depleted SOI technology chip die photo.

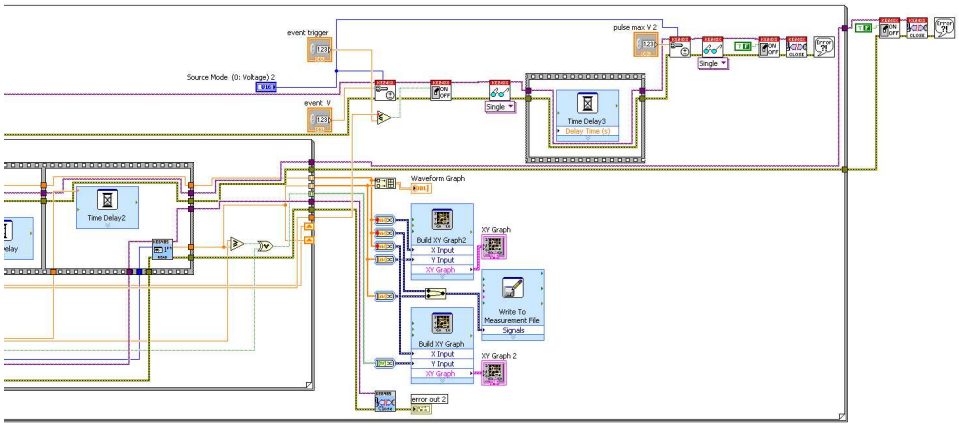
Appendix B

Chip design

B.0.3 Labview block diagram



(a) MITLL FDSOI 150nm PMOS Vt mismatch block diagram (part 1).



(b) MITLL FDSOI 150nm PMOS Vt mismatch block diagram (part 2).

Figure B.1: Automated test setup block diagram for sequentially measuring multiple PMOS devices by decode gate selection.

Appendix C

Analytical derivation of read delay as a function of Vwl, Vddc, and Vtn$_0$

I_{pd} is in the linear mode, therefore:

$$I_{pd}(Va, Vb) = k_n \cdot \frac{W_{pd}}{L_{pd}} \cdot (Va - Vtn_0 - \frac{Vb}{2}) \tag{C.1}$$

Where k_n is the product of mobility and Cox, Va is the gate voltage supplied by the latch cross couple. Vb is the internal voltage determined by the voltage divider relationship between the PG and PD NMOS devices. Vtn_0 is the NMOS threshold voltage with no body effect. I_{pg} in velocity saturation:

I_{pg} is in the linear mode, therefore:

$$I_{pg}(Vb) = k_n \cdot \frac{W_{pg}}{L_{pg}} \cdot \Psi_n \cdot (Vdd - Vb - Vtn_0 + \frac{\gamma_n \cdot Vb}{2} - \frac{\Psi_n}{2}) \cdot [1 + \lambda_n \cdot (Vdd - Vb)] \tag{C.2}$$

where Ψ_n is the NMOS velocity saturation voltage, and λ_n is the NMOS channel length modulation effect.

Simplification 1: neglect channel length modulation

Simplification 2: linearize body effect in the following way:

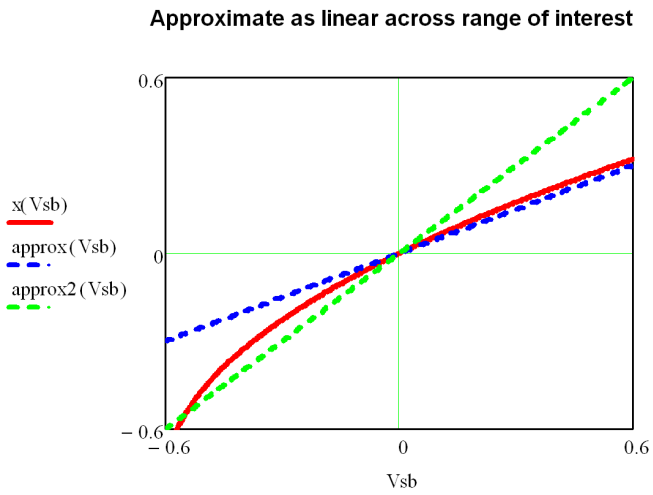


Figure C.1: A linear approximation used for NMOS body effect across the range of interest for a tractable algebraic solution.

Simplification 3: neglect small Vb/2 term in linear model

This allows the voltage on node b (Vnb) to be expressed as:

$$Vnb(Vwl, Vddc, Vtn_0) = \frac{\Psi_n \cdot Vwl - \Psi_n \cdot Vtn_0 - \frac{\Psi_n^2}{2}}{\beta \cdot Vddc - \beta \cdot Vtn_0 + \Psi_n - \Psi_n \cdot \frac{\lambda_n}{\gamma_x}} \tag{C.3}$$

By substitution the read current can be written as:

$$\begin{aligned} Iread(Vwl, Vddc, Vtn_0) = k_n \cdot \frac{W_{pg}}{L_{pg}} \cdot \Psi_n (Vwl - Vnb(Vwl, Vddc, Vtn_0) \\ -Vtn_0 + \frac{\gamma_n \cdot Vnb(Vwl, Vddc, Vtn_0)}{\gamma_x} - \frac{\Psi_n}{2}) \end{aligned} \tag{C.4}$$

The read delay is expressed as:

$$\tau_{read}(Vwl, Vddc, Vtn_0) = \frac{C_{BL} \cdot \delta V_{BL}}{Iread(Vwl, Vddc, Vtn_0)} \tag{C.5}$$

Where CBL is the BL capacitance and VBL is the delta voltage that must be developed on the BL to successfully read. 100mV is used in this calculation.

The read delay as s function of Vtn, Vwl and Vddc is then expressed as:

$$\tau_{read}(Vwl, Vddc, Vtn_0) = \frac{2C_{BL} \cdot L_{pg}\delta V_{BL}(\Psi_n\gamma_x - \Psi_n\gamma_n - Vtn_0\beta\gamma_x + Vddc\beta\gamma_x)}{W_{pg}\Psi_n\beta k_n\gamma_x(Vtn_0 - Vddc) \cdot (2Vtn_0 - 2Vwl + \Psi_n)} \tag{C.6}$$

Appendix D

Publications related to this thesis

D.0.4 Related Publications

1. R. W. Mann, L. A. Clevenger and Q. Z. Hong, “The C49 to C54-TiSi2 transformation in self-aligned silicide C54-TiSi2 transformation in self-aligned silicide applications,” J. Appl. Phys. 73 (7) p.3566-3568 (1993).

2. C. Koburger III, W. Clark, J. Adkisson, E. Adler, P. Bakeman, A. Bergendahl, A. Botula, W. Chang, B. Davari, J. Givens, H. Hansen, S. Holmes, D. Horak, C. Lam, J. Lasky, S. Luce, R. Mann, G. Miles, J. Nakos, E. Nowak, G. Shahidi, Y. Taur, F. White, and M. Wordeman, “A half micron CMOS logic generation,” IBM J. Res. Dev. (USA) Vol.39, No.1-2 Jan. March 1995, p.215-27

3. R.W.Mann, L.A.Clevenger, G.L.Miles, J.M.E.Harper, F.M.D’Heurle and C.Cabral,Jr., T.A.Knotts, D.W.Rakowski, “Reduction of the C54-TiSi2 phase transformation temperature using refractory metal ion implantation,” Applied Physics Letters, vol. 67, no. 25, pp. 3729-3731, 1995.

4. R.W. Mann, L. Clevenger, P. Agnello, and F. White, “Silicides and local interconnects for high-performance VLSI applications: a review,” IBM Journal of Res. and Dev., Vol.39, No.4 July P403-17 (1995).

5. P. Smeys, V. McGahay, I. Yang, J. Adkisson, K. Beyer, O. Bula, Z. Chen, B. Chu, J. Culp, S. Das, A. Eckert, L. Hadel, M. Hargrove, J. Herman, L. Lin, R. Mann, E. Maciejewski, S. Narasimha, P. O’Neil, S. Rauch, D. Ryan, J. Toomey, L. Tsou, P. Varekamp, R. Wachnik, T. Wagner, S. Wu, C. Yu, P. Agnello, J. Connolly, S. Crowder, C. Davis, R. Ferguson, A. Sekiguchi, L. Su, R. Goldblatt, and T. C. Chen, “0.13 m high performance SOI technology development,” VLSI Tech Symp paper 19.1 (2000).

6. S. V. Kosonocky, M, Immediato, P. Cottrell, T. Hook, R. Mann, J. Brown, “Enhanced Mult-Threshold (MTCMOS) Circuits Using Variable Well Bias, Low Power Electronics and Design,” International Symposium on, 2001. 6-7 Aug. 2001 Page(s):165 - 169

7. Z. Luo, A. Steegen, M. Eller, R. Mann, C. Baiocco, P. Nguyen, L. Kim, M. Hoinkis, V. Ku, V. Klee, F. Jamin, P. Wrschka, P. Shafer, W. Lin, S. Fang, W. Tan, D. Park, R. Mo, J. Lian, D. Vietzke, C. Coppock, A. Vayshenker, T. Hook, V. Chan, K. Kim, A. Cowley, S. Kim, E. Kaltalioglu, B. Zhang, S. Marokkey, Y. Lin, M. Weybright, R. Rengarajan, J. Ku, T. Schiml, J. Sudijono, I. Yang, Clement Wann, “High Performance and Low Power Transistors Integrated in 65nm Bulk CMOS Technology,” Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International 13-15 Dec. 2004 Page(s):661 - 664

8. A. Steegen, R. Mo, R. Mann, M. C. Sun, M. Eller, G. Leake, D. Vietzke, A. Tilke, F. Guarin, A. Fischer, T. Pompl, G. Massey, A. Vayshenker, W. L. Tan, A. Ebert, W. Lin, W. Gao, J. Lian, J. P. Kim, P. Wrschka, J. H. Yang, A. Ajmera, R. Knoefler, Y. W. Teh, F. Jamin, J. E. Park, K. Hooper, C. Griffin, P. Nguyen, V. Klee, V. Ku, C. Baiocco, G. Johnson, L. Tai, J. Benedict, S. Scheer, H. Zhuang, V. Ramanchandran, G. Matusiewicz, Y. H. Lin, Y. K. Siew, F. Zhang, L. S. Leong, S. L. Liew, K. C. Park, K. W. Lee, D. H. Hong, S. M. Choi, E. Kaltalioglu, S. O. Kim, M. Naujok, M. Sherony, A. Cowley, A. Thomas, J. Sudijohno, T. Schiml, J. H. Ku, and I. Yang, "65nm CMOS technology for low power applications," Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International 5-7 Dec. 2005 Page(s):64 - 67

9. R. W. Mann, W. Abadeer, M. Brietwisch, O. Bula, J. Brown, B. Colwill, P.Cottrell, W. Crocco, S. Furkay, M. Hauser, T. Hook, D. Hoyniak, J. Johnson, C. Lam, R. Mih, J. Rivard, A. Moriwaki, E. Phipps, C. Putnam, B. Rainey, J. Toomey, M. Younus, "Ultralow-Power SRAM technology," IBM J. Res. Dev. Vol. 47, no. 5/6 Sep/Nov 2003, p. 553-566

10. Benton H. Calhoun, Sudhanshu Khanna, Randy Mann, and Jiajing Wang,"Sub-threshold Circuit Design with Shrinking CMOS Devices," International Symposium on Circuits and Systems, 2009.

11. C. Wann, R. Wong, D. J. Frank, 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, VLSI Technology," 2005. (VLSI-TSA-Tech). 2005 IEEE VLSI-TSA International Symposium on 25-27 April 2005 Page(s):21 - 22

12. T. B. Hook, M. Breitwisch, J. Brown, P. Cottrell, D. Hoyniak, C. Lam, and R. Mann, “Noise Margin and Leakage in Ultra-Low Leakage SRAM Cell Design,” IEEE Trans. Elec. Dev. Vol. 49, no. 8, Aug. 2002, p. 1499-1501

13. D. Cole, O. Bula, E. Conrad, D. Coops, W. Leipold, R. Mann, and J. Oppold, “Optimization Criteria for SRAM Design- Lithography Contribution,” Proceedings of SPIE - The International Society for Optical Engineering v 3679 n II 1999. p 847-859.

14. Benton H. Calhoun, Sudhanshu Khanna, Randy Mann, and Jiajing Wang. Subthreshold circuit design with shrinking CMOS devices. In Proc. IEEE International Symposium on Circuits and Systems ISCAS 2009, pages 25412544, May 24 2009Yearly 27 2009.

15. R.W.Mann, S.Nalam, S.Khanna, J.Wang, G.Braceras, H.Pilo, B.H.Calhoun, “Impact of circuit assist methods on margin and performance in 6T SRAM,” Solid-State Electronics, 54(11):13981407, Nov 2010.

16. R. W. Mann, S. Nalam, J. Wang, B. H. Calhoun, “Limits of Bias Based Assist Methods in Nano-Scale 6T SRAM,” in Proc. 11th International Symposium on Quality Electronic Design ISQED ’10’ 2010 pp. 1-6

17. Jiajing Wang, Satyanand Nalam, Zhenyu(Jerry) Qi, Randy W. Mann, Mircea Stan, and Benton H. Calhoun, Improving SRAM Vmin and Yield by Using Variation-Aware BTI Stress, CICC, San Jose, CA, 09/2010.

Appendix E

Patents related to this thesis

E.0.5 Related Patents

1. 7,087,486 Method for scalable, low-cost polysilicon capacitor in a planar DRAM
2. 7,057,180 Detector for alpha particle or cosmic ray
3. 6,489,223 Angled implant process
4. 6,187,679 Low temperature formation of low resistivity titanium silicide
5. 7,005,334 Zero threshold voltage pFET and method of making same
6. 6,144,086 Structure for improved latch-up using dual depth STI with impurity implant
7. 6,962,838 High mobility transistors in SOI and method for forming
8. 6,946,376 Symmetric device with contacts self aligned to gate
9. 6,614,124 Simple 4T static ram cell for low power CMOS applications

10. 6,420,746 Three device DRAM cell with integrated capacitor and local interconnect
11. 6,967,351 Finfet SRAM cell using low mobility plane for cell stability and method for forming
12. 6,778,449 Method and design for measuring SRAM array leakage macro (ALM)
13. 7,313,032 SRAM voltage control for improved operational margins
14. 7,075,153 Grounded body SOI SRAM cell
15. U.S. Provisional Patent Application Serial No. 61/365,962 Improved Dense 6T SRAM Cell Layout Structure and Related Method

Bibliography

[1] M. H. Abu-Rahma, M. Anis, and Sei Seung Yoon. A robust single supply voltage SRAM read assist technique using selective precharge. In *Proc. 34th European Solid-State Circuits Conference ESSCIRC 2008*, pages 234–237, 15–19 Sept. 2008.

[2] K. Agarwal and S. Nassif. The impact of random device variation on sram cell stability in sub-90-nm cmos technologies. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(1):86 –97, jan. 2008.

[3] 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. pages 655 –658, dec. 2005.

[4] R. E. Aly and M. A. Bayoumi. Low-power cache design using 7t sram cell. *IEEE Trans. Circuits Syst. II*, 54(4):318–322, April 2007.

[5] A. Asenov, A.R. Brown, J.H. Davies, S. Kaya, and G. Slavcheva. Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs. *IEEE Transactions on Electron Devices*, 50(9):1837–1852, 2003.

[6] A. Asenov and S. Saini. Suppression of random dopant-induced threshold voltage fluctuations in sub-0.1-& mosfet's with epitaxial and &doped channels. *IEEE Trans. Electron Devices*, 46(8):1718–1724, Aug. 1999.

[7] R. Baumann. The impact of technology scaling on soft error rate performance and limits to the efficacy of error correction. In *Proc. Digest. International Electron Devices Meeting IEDM '02*, pages 329–332, 8–11 Dec. 2002.

[8] A. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Qiuyi Ye, and Ken Chin. Fluctuation limits & scaling opportunities for CMOS SRAM cells. In *Proc. IEDM Technical Digest Electron Devices Meeting IEEE International*, pages 659–662, 5–5 Dec. 2005.

[9] A. J. Bhavnagarwala, S. Kosonocky, C. Radens, Yuen Chan, K. Stawiasz, U. Srinivasan, S. P. Kowalczyk, and M. M. Ziegler. A sub-600-mv, fluctuation tolerant 65-nm CMOS SRAM array with dynamic cell biasing. *IEEE J. Solid-State Circuits*, 43(4):946–955, April 2008.

[10] S. Bordez, A. Cathignol, and K. Rochereau. A continuous model for MOSFET vt matching considering additional length effects. In *Proc. IEEE International Conference on Microelectronic Test Structures ICMTS '07*, pages 226–229, 19–22 March 2007.

[11] B. H. Calhoun and A. Chandrakasan. A 256kb sub-threshold SRAM in 65nm CMOS. In *Proc. Digest of Technical Papers. IEEE International Solid-State Circuits Conference ISSCC 2006*, pages 2592–2601, 6–9 Feb. 2006.

[12] Benton H. Calhoun, Sudhanshu Khanna, Randy Mann, and Jiajing Wang. Sub-threshold circuit design with shrinking CMOS devices. In *Proc. IEEE International Symposium on Circuits and Systems ISCAS 2009*, pages 2541–2544, May 24 2009–Yearly 27 2009.

[13] B.H. Calhoun and A. Chandrakasan. Analyzing static noise margin for sub-threshold SRAM in 65nm CMOS. pages 363 – 366, sep. 2005.

[14] B.H. Calhoun and A.P. Chandrakasan. Static noise margin variation for sub-threshold SRAM in 65-nm CMOS. *Solid-State Circuits, IEEE Journal of*, 41(7):1673 –1679, jul. 2006.

[15] E.H. Cannon, D.D. Reinhardt, M.S. Gordon, and P.S. Makowenskyj. SRAM SER in 90, 130 and 180 nm bulk and SOI technologies. pages 300 – 304, apr. 2004.

[16] 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. In *Proc. Digest of Technical Papers VLSI Technology 2005 Symposium on*, pages 128–129, 14–16 June 2005.

[17] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek. An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches. *IEEE J. Solid-State Circuits*, 43(4):956–963, April 2008.

[18] Y. H. Chen, W. M. Chan, S. Y. Chou, H. J. Liao, H. Y. Pan, J. J. Wu, C. H. Lee, S. M. Yang, Y. C. Liu, and H. Yamauchi. A 0.6V 45nm adaptive dual-rail SRAM compiler circuit design for lower VDDmin VLSIs. In *Proc. IEEE Symposium on VLSI Circuits*, pages 210–211, 18–20 June 2008.

[19] Yeonbae Chung and Seung-Ho Song. Implementation of low-voltage static RAM with enhance data stability and circuit speed. *Microelectronics Journal*, 40:944–951, 2009.

[20] R. Difrenza, K. Rochereau, T. Devoivre, B. Tavel, B. Duriez, D. Roy, S. Jullian, A. Dezzani, R. Boulestin, P. Stolk, and F. Arnaud. MOSFET matching improvement in 65nm technology providing gain on both analog and SRAM performances. pages 137 – 142, apr. 2005.

[21] P. E. Dodd and L. W. Massengill. Basic mechanisms and modeling of single-event upset in digital microelectronics. *IEEE Trans. Nucl. Sci.*, 50(3):583–602, June 2003.

[22] W.F. Ellis, R.W. Mann, D.J. Wager, and R. C. Wong. SRAM voltage control for improved operational margins, 2007.

[23] G. Georgakos, P. Huber, M. Ostermayr, E. Amirante, and F. Ruckerbauer. Investigation of Increased Multi-Bit Failure Rate Due to Neutron Induced SEU in Advanced Embedded SRAMs. pages 80–81, 2007.

[24] G. Ghibaudo and T. Boutchacha. Electrical noise and RTS fluctuations in advanced CMOS devices. *Microelectronics Reliability*, 42(4-5):573 – 582, 2002.

[25] N. Gierczynski, B. Borot, N. Planes, and H. Brut. A new combined methodology for write-margin extraction of advanced SRAM. In *IEEE International Conference on Microelectronic Test Structures, 2007. ICMTS'07*, pages 97–100, 2007.

[26] B.S. Haran, A. Kumar, L. Adam, J. Chang, V. Basker, S. Kanakasabapathy, D. Horak, S. Fan, J. Chen, J. Faltermeier, S. Seo, M. Burkhardt, S. Burns, S. Halle, S. Holmes, R. Johnson, E. McLellan, T.M. Levin, Y. Zhu, J. Kuss, A. Ebert, J. Cummings, D. Canaperi, S. Paparao, J. Arnold, T. Sparks, C.S. Koay, T. Kanarsky, S. Schmitz, K. Petrillo, R.H. Kim, J. Demarest, L.F. Edge, H. Jagannathan, M. Smalley, N. Berliner, K. Cheng, D. LaTulipe, C. Koburger, S. Mehta, M. Raymond, M. Colburn, T. Spooner, V. Paruchuri, W. Haensch, D. McHerron, and B. Doris. 22 nm technology compatible fully functional 0.1 μm^2 6T SRAM cell. In *Proc. IEEE International Electron Devices Meeting IEDM 2008*, pages 1–4, 15-17 Dec. 2008.

[27] P. Hazucha and C. Svensson. Impact of CMOS technology scaling on the atmospheric neutron soft error rate. *IEEE Trans. Nucl. Sci.*, 47(6):2586–2594, Dec. 2000.

[28] T. Heijmen, B. Kruseman, R. van Veen, and M. Meijer. Technology scaling of critical charges in storage circuits based on cross-coupled inverter-pairs. pages 675 – 676, apr. 2004.

[29] O. Hirabayashi, A. Kawasumi, A. Suzuki, Y. Takeyama, K. Kushida, T. Sasaki, A. Katayama, G. Fukano, Y. Fujimura, T. Nakazato, Y. Shizuki, N. Kushiyama, and T. Yabe. A process-variation-tolerant dual-power-supply SRAM with

0.179μm^2 cell in 40nm CMOS using level-programmable wordline driver. In *Proc. IEEE International Solid-State Circuits Conference - Digest of Technical Papers ISSCC 2009*, pages 458–459,459a, 8–12 Feb. 2009.

[30] T. B. Hook, J. Brown, P. Cottrell, E. Adler, D. Hoyniak, J. Johnson, and R. Mann. Lateral ion implant straggle and mask proximity effect. *IEEE Trans. Electron Devices*, 50(9):1946–1951, Sept. 2003.

[31] T.B. Hook, M. Breitwisch, J. Brown, P. Cottrell, D. Hoyniak, Chung Lam, and R. Mann. Noise margin and leakage in ultra-low leakage SRAM cell design. *Electron Devices, IEEE Transactions on*, 49(8):1499 – 1501, aug. 2002.

[32] T.B. Hook, J.S. Brown, M. Breitwisch, D. Hoyniak, and R. Mann. High-performance logic and high-gain analog CMOS transistors formed by a shadow-mask technique with a single implant step. *IEEE Transactions on Electron Devices*, 49(9):1623–1627, 2002.

[33] T.B. Hook, J.B. Johnson, J-P. Han, A. Pond, T. Shimiu, and G. Tsutsui. Channel length and threshold voltage dependence of transistor mismatch in a 32nm hkmg technology. *IEEE Transactions on Electron Devices*, to be published.

[34] T.B Hook and G. Leak. Method of selectively adjusting ion implantation dose on semiconductor devices, 2010.

[35] M. Iijima, K. Seto, M. Numa, A. Tada, and T. Ipposhi. Low power SRAM with boost driver generating pulsed word line voltage for sub-1V operation. *JCP 3*, 5:34–40, 2008.

[36] S. Inouye, M. Robles-Bruce, and M. Scherer. 2010 microprocessors. http://www.databeans.net/reports, March 2010.

[37] M. Ishida, T. Kawakami, A. Tsuji, N. Kawamoto, M. Motoyoshi, and N. Ouchi. A novel 6T-SRAM cell technology designed with rectangular patterns scalable beyond 0.18 generation and desirable for ultra high speed operation. In *Proc.*

International Electron Devices Meeting IEDM '98 Technical Digest, pages 201–204, 6–9 Dec. 1998.

[38] T. Jhaveri, V. Rovner, L. Liebmann, L. Pileggi, A.J. Strojwas, and J.D. Hibbeler. Co-optimization of circuits, layout and lithography for predictive technology scaling beyond gratings. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 29(4):509 –527, april 2010.

[39] J. B. Johnson, T. B. Hook, and Yoo-Mi Lee. Analysis and modeling of threshold voltage mismatch for CMOS at 65 nm and beyond. *IEEE Electron Device Lett.*, 29(7):802–804, July 2008.

[40] K. Kang, H. Kufluoglu, K. Roy, and M. Ashraful Alam. Impact of negative-bias temperature instability in nanoscale SRAM array: Modeling and analysis. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 26(10):1770–1781, Oct. 2007.

[41] R.W. Keyes. The effect of randomness in the distribution of impurity atoms on FET thresholds. *Applied Physics A: Materials Science & Processing*, 8(3):251–259, 1975.

[42] M. Khellah, Nam Sung Kim, Yibin Ye, D. Somasekhar, T. Karnik, N. Borkar, F. Hamzaoglu, T. Coan, Yih Wang, Kevin Zhang, C. Webb, and Vivek De. PVT-variations and supply-noise tolerant 45nm dense cache arrays with diffusion-notch-free (dnf) 6T SRAM cells and dynamic multi-Vcc circuits. In *Proc. IEEE Symposium on VLSI Circuits*, pages 48–49, 18–20 June 2008.

[43] M. Khellah, Nam Sung Kim, Yibin Ye, D. Somasekhar, T. Karnik, N. Borkar, G. Pandya, F. Hamzaoglu, T. Coan, Yih Wang, K. Zhang, C. Webb, and V. De. Process, temperature, and supply-noise tolerant 45formula formulatype="inline"¿ $/formula > nm$ dense cache arrays with diffusion-notch-free (dnf) 6T SRAM cells and dynamic multi-Vcc circuits. *IEEE J. Solid-State Circuits*, 44(4):1199–1208, April 2009.

[44] M. Khellah, Yibin Ye, Nam Sung Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De. Wordline & bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65nm CMOS designs. In *Proc. Digest of Technical Papers VLSI Circuits 2006 Symposium on*, pages 9–10, 2006.

[45] A. T. Krishnan, V. Reddy, D. Aldrich, J. Raval, K. Christensen, J. Rosal, C. O'Brien, R. Khamankar, A. Marshall, W. K. Loh, R. McKee, and S. Krishnan. SRAM cell static noise margin and Vmin sensitivity to transistor degradation. In *Proc. International Electron Devices Meeting IEDM '06*, pages 1–4, 11–13 Dec. 2006.

[46] K. Kushida, A. Suzuki, G. Fukano, A. Kawasumi, O. Hirabayashi, Y. Takeyama, T. Sasaki, A. Katayama, Y. Fujimura, and T. Yabe. A 0.7V single-supply SRAM with 0.495 um^2 cell in 65nm technology utilizing self-write-back sense amplifier and cascaded bit line scheme. In *Proc. IEEE Symposium on VLSI Circuits*, pages 46–47, 18–20 June 2008.

[47] Jonghwan Lee, G. Bosman, K.R. Green, and D. Ladwig. Model and analysis of gate leakage current in ultrathin nitrided oxide MOSFETs. *Electron Devices, IEEE Transactions on*, 49(7):1232 –1241, jul. 2002.

[48] Ying Li, Guofu Niu, J.D. Cressler, J. Patel, C.J. Marshall, P.W. Marshall, H.S. Kim, R.A. Reed, and M.J. Palmer. Anomalous radiation effects in fully depleted SOI MOSFETs fabricated on SIMOX. *Nuclear Science, IEEE Transactions on*, 48(6):2146 –2151, dec. 2001.

[49] L. Liebmann, L. Pileggi, J. Hibbeler, V. Rovner, T. Jhaveri, and G. Northrop. Simplify to survive, prescriptive layouts ensure profitable scaling to 32nm and beyond. 2009.

[50] Sheng Lin, Yong-Bin Kim, and F. Lombardi. A 32nm SRAM design for low power and high stability. In *Proc. 51st Midwest Symposium on Circuits and Systems MWSCAS 2008*, pages 422–425, 10–13 Aug. 2008.

[51] Zhiyu Liu and V. Kursun. Characterization of a novel nine-transistor SRAM cell. *IEEE Trans. VLSI Syst.*, 16(4):488–492, April 2008.

[52] S.-H. Lo, D.A. Buchanan, Y. Taur, and W. Wang. Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's. *Electron Device Letters, IEEE*, 18(5):209 –211, may. 1997.

[53] DL Losee, JP Lavine, EA Trabka, S.T. Lee, and CM Jarman. Phosphorus diffusion in polycrystalline silicon. *Journal of Applied Physics*, 55:1218, 1984.

[54] Z. Luo, N. Rovedo, S. Ong, B. Phoong, M. Eller, H. Utomo, C. Ryou, H. Wang, R. Stierstorfer, L. Clevenger, S. Kim, J. Toomey, D. Sciacca, J. Li, W. Wille, L. Zhao, L. Teo, T. Dyer, S. Fang, J. Yan, O. Kwon, D. Park, J. Holt, J. Han, V. Chan, T. K. J. Yuan, H. Lee, S. Lee, A. Vayshenker, Z. Yang, C. Tian, H. Ng, H. Shang, M. Hierlemann, J. Ku, J. Sudijono, and M. Ieong. High performance transistors featured in an aggressively scaled 45nm bulk CMOS technology. In *Proc. IEEE Symposium on VLSI Technology*, pages 16–17, 12–14 June 2007.

[55] Z. Luo, A. Steegen, M. Eller, R. Mann, C. Baiocco, P. Nguyen, L. Kim, M. Hoinkis, V. Ku, V. Klee, F. Jamin, P. Wrschka, P. Shafer, W. Lin, S. Fang, A. Ajmera, W. Tan, D. Park, R. Mo, J. Lian, D. Vietzke, C. Coppock, A. Vayshenker, T. Hook, V. Chan, K. Kim, A. Cowley, S. Kim, E. Kaltalioglu, B. Zhang, S. Marokkey, Y. Lin, K. Lee, H. Zhu, M. Weybright, R. Rengarajan, J. Ku, T. Schiml, J. Sudijono, I. Yang, and C. Wann. High performance and low power transistors integrated in 65nm bulk CMOS technology. In *Proc. IEDM Technical Digest Electron Devices Meeting IEEE International*, pages 661–664, 2004.

[56] Randy W. Mann, Satyanand Nalam, Jiajing Wang, and Benton H. Calhoun. Limits of bias based assist methods in nano-scale 6T SRAM. In *Proc. 11th International Symposium on Quality Electronic Design, ISQED*, March 2010.

[57] Randy W. Mann, Jiajing Wang, Satyanand Nalam, Sudhanshu Khanna, Geordie Braceras, Harold Pilo, and Benton H. Calhoun. Impact of circuit assist methods

on margin and performance in 6T SRAM. *Solid-State Electronics*, 54(11):1398 – 1407, Nov 2010.

[58] R.W. Mann, W. Abadeer, M.J. Breitwisch, O. Bula, J.S. Brown, B.C. Colwill, P.E. Cottrell, W.G. Crocco, S. Furkay, M.J. Hauser, T. Hook, D. Hoyniak, J. Johnson, C. Lam, R. Mih, J. Rivard, A. Moriwaki, E. Phipps, C. Putnam, B. Rainey, J. Toomey, and M. Younus. Ultralow-power SRAM technology. *IBM J. Res. & Dev.*, Vol. 47:553–566, 2003.

[59] R.W. Mann and B. H. Calhoun. Improved dense 6T SRAM cell layout structure and related method, 2010.

[60] J. Mc Ginley, O. Noblanc, C. Julien, S. Parihar, K. Rochereau, R. Difrenza, and P. Llinares. Impact of pocket implant on MOSFET mismatch for advanced CMOS technology. In *Proc. International Conference on Microelectronic Test Structures ICMTS '04*, pages 123–126, 22–25 March 2004.

[61] B. Mohammad, M. Saint-Laurent, P. Bassett, and J. Abraham. Cache design for low power and high yield. In *Proc. 9th International Symposium on Quality Electronic Design ISQED 2008*, pages 103–107, 17–19 March 2008.

[62] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Reduction of parametric failures in sub-100-nm SRAM array using body bias. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 27(1):174–183, Jan. 2008.

[63] S. Nalam and B. H. Calhoun. Asymmetric sizing in a 45nm 5T SRAM to improve read stability over 6T. In *Proc. IEEE Custom Integrated Circuits Conference CICC '09*, pages 709–712, 13–16 Sept. 2009.

[64] K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, Y. Oda, K. Usui, T. Kawamura, N. Tsuboi, T. Iwasaki, K. Hashimoto, H. Makino, and H. Shinohara. A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment. In *Proc. IEEE Symposium on VLSI Circuits*, pages 212–213, 18–20 June 2008.

[65] H. Noguchi, S. Okumura, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, and M. Yoshimoto. Which is the best dual-port sram in 45-nm process technology? 8T, 10T single end, and 10T differential . In *Proc. IEEE International Conference on Integrated Circuit Design and Technology and Tutorial ICICDT 2008*, pages 55–58, 2–4 June 2008.

[66] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, and H. Shinohara. A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits. *IEEE J. Solid-State Circuits*, 42(4):820–829, April 2007.

[67] S. Okumura, Y. Iguchi, S. Yoshimoto, H. Fujiwara, H. Noguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto. A 0.56-V 128kb 10T SRAM using column line assist (cla) scheme. In *Proc. Quality of Electronic Design Quality of Electronic Design ISQED 2009*, pages 659–663, 16–18 March 2009.

[68] A. Papoulis and S.U. Pillai. *Probability, random variables, and stochastic processes*. McGraw-Hill New York, 2002.

[69] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers. Matching properties of mos transistors. *IEEE J. Solid-State Circuits*, 24(5):1433–1439, Oct 1989.

[70] 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. In *Proc. Digest of Technical Papers VLSI Circuits 2006 Symposium on*, pages 15–16, 2006.

[71] I. Polishchuk, N. Mathur, C. Sandstrom, P. Manos, and O. Pohland. CMOS Vt-control improvement through implant lateral scatter elimination. In *IEEE International Symposium on Semiconductor Manufacturing, 2005. ISSM 2005*, pages 193–196, 2005.

[72] H. Puchner and S. Selberherr. An advanced model for dopant diffusion in polysilicon. *IEEE Transactions on Electron Devices*, 42(10):1750–1755, 1995.

[73] K. Samsudin, F. Adamu-Lema, A.R. Brown, S. Roy, and A. Asenov. Combined sources of intrinsic parameter fluctuations in sub-25nm generation UTB-SOI MOSFETs: A statistical simulation study. *Solid-State Electronics*, 51(4):611 – 616, 2007. Special Issue: Papers selected from the 2006 ULIS Conference.

[74] E. Seevinck, FJ. List, and J. Lohstroh. Static-noise margin analysis of MOS SRAM cells. *IEEE Journal of Solid-State Circuits*, SC-22:748–754, 1987.

[75] N. Seifert, P. Slankard, M. Kirsch, B. Narasimham, V. Zia, C. Brookreson, A. Vo, S. Mitra, B. Gill, and J. Maiz. Radiation-induced soft error rates of advanced CMOS bulk devices. In *Proc. 44th Annual. IEEE International Reliability Physics Symposium*, pages 217–225, 26–30 March 2006.

[76] Y.M. Sheu, K.W. Su, S. Tian, S.J. Yang, C.C. Wang, M.J. Chen, and S. Liu. Modeling the well-edge proximity effect in highly scaled MOSFETs. *IEEE Transactions on Electron Devices*, 53(11):2792–2798, 2006.

[77] N. Shibata, H. Kiya, S. Kurita, H. Okamoto, M. Tan'no, and T. Douseki. A 0.5-V 25-MHz 1-mW 256-kb MTCMOS/SOI SRAM for solar-power-operated portable personal digital equipment - sure write operation by using step-down negatively overdriven bitline scheme. *IEEE J. Solid-State Circuits*, 41(3):728–742, March 2006.

[78] J. Shimokawa, M. Sato, C. Suzuki, M. Nakamura, and Y. Ohji. Theoretical approach and precise description of PBTI in high-k gate dielectrics based on electron trap in pre-existing and stress-induced defects. pages 973 –976, apr. 2009.

[79] T. Suzuki, H. Yamauchi, Y. Yamagami, K. Satomi, and H. Akamatsu. A stable 2-port SRAM cell design against simultaneously read/write-disturbed accesses. *IEEE J. Solid-State Circuits*, 43(9):2109–2119, Sept. 2008.

[80] S.M. Sze. *VLSI technology*. McGraw-Hill New York, 1988.

[81] Y. Taur and T.H. Ning. *Fundamentals of Modern VLSI Devices*. Cambridge Univ. Press, 1998.

[82] Seng Oon Toh, Y. Tsukamoto, Zheng Guo, L. Jones, Tsu-Jae King Liu, and B. Nikolic. Impact of random telegraph signals on vmin in 45nm SRAM. pages 1 –4, dec. 2009.

[83] B. Wang, A. Norwood, and J. Unsworth. Gartner research press release. http://www.gartner.com, Sept 2010.

[84] D. P. Wang, H. J. Liao, H. Yamauchi, Y. H. Chen, Y. L. Lin, S. H. Lin, D. C. Liu, H. C. Chang, and W. Hwang. A 45nm dual-port SRAM with write and read capability enhancement at low voltage. In *Proc. IEEE International SOC Conference*, pages 211–214, 26–29 Sept. 2007.

[85] Jiajing Wang, S. Nalam, and B.H. Calhoun. Analyzing static and dynamic write margin for nanometer SRAMs. pages 129 –134, aug. 2008.

[86] Jiajing Wang, Satyanand Nalam, and Benton H. Calhoun. Analyzing static and dynamic write margin for nanometer SRAMs. In *Proc. Int. Symp. on Low power electronics and design*, pages 129–134, New York, NY, USA, 2008. ACM.

[87] C. C. Wu, Y. K. Leung, C. S. Chang, M. H. Tsai, H. T. Huang, D. W. Lin, Y. M. Sheu, C. H. Hsieh, W. J. Liang, L. K. Han, W. M. Chen, S. Z. Chang, S. Y. Wu, S. S. Lin, H. C. Lin, C. H. Wang, P. W. Wang, T. L. Lee, C. Y. Fu, C. W. Chang, S. C. Chen, S. M. Jang, S. L. Shue, H. T. Lin, Y. C. See, Y. J. Mii, C. H. Diaz, B. J. Lin, M. S. Liang, and Y. C. Sun. A 90-nm CMOS device technology with high-speed, general-purpose, and low-leakage transistors for system on chip applications. In *Proc. Digest. International Electron Devices Meeting IEDM '02*, pages 65–68, 8–11 Dec. 2002.

[88] Shien-Yang 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, H. F. Chen, S. Y. Chang, S. H. Wang, P. Y.

Tong, Y. L. Hsieh, 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, K. C. Lin, L. Y. Yeh, K. C. Ku, S. C. Chen, 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. In *Proc. IEEE International Electron Devices Meeting IEDM 2007*, pages 263–266, 10–12 Dec. 2007.

[89] Shien-Yang Wu, J.J. Liaw, C.Y. Lin, M.C. Chiang, C.K. Yang, J.Y. Cheng, M.H. Tsai, M.Y. Liu, P.H. Wu, C.H. Chang, L.C. Hu, C.I. Lin, H.F. Chen, S.Y. Chang, S.H. Wang, P.Y. Tong, Y.L. Hsieh, K.H. Pan, C.H. Hsieh, C.H. Chen, C.H. Yao, C.C. Chen, T.L. Lee, C.W. Chang, H.J. Lin, S.C. Chen, J.H. Shieh, S.M. Jang, K.S. Chen, Y. Ku, Y.C. See, and W.J. Lo. A highly manufacturable 28nm CMOS low power platform technology with fully functional 64Mb SRAM using dual/tripe gate oxide process. In *VLSI Technology, 2009 Symposium on*, pages 210 –211, 16-18 2009.

[90] M. Yamaoka, N. Maeda, Y. Shimazaki, and K. Osada. 65nm low-power high-density SRAM operable at 1.0V under 3σ systematic variation using separate Vth monitoring and body bias for NMOS and PMOS. In *Proc. Digest of Technical Papers. IEEE International Solid-State Circuits Conference ISSCC 2008*, pages 384–622, 3–7 Feb. 2008.

[91] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara. Low-power embedded SRAM modules with expanded margins for writing. In *Proc. Digest of Technical Papers Solid-State Circuits Conference ISSCC. 2005 IEEE International*, pages 480–611, 10–10 Feb. 2005.

[92] M. Yamaoka, K. Osada, and K. Ishibashi. 0.4-V logic-library-friendly SRAM array using rectangular-diffusion cell and delta-boosted-array voltage scheme. *IEEE J. Solid-State Circuits*, 39(6):934–940, June 2004.

[93] H. Yamauchi. Embedded SRAM circuit design technologies for a 45nm and beyond. pages 1028 –1033, oct. 2007.

[94] H. S. Yang, R. Wong, R. Hasumi, Y. Gao, N. S. Kim, D. H. Lee, S. Badrudduza, D. Nair, M. Ostermayr, H. Kang, H. Zhuang, J. Li, L. Kang, X. Chen, A. Thean, F. Arnaud, L. Zhuang, C. Schiller, D. P. Sun, Y. W. Teh, J. Wallner, Y. Takasu, K. Stein, S. Samavedam, D. Jaeger, C. V. Baiocco, M. Sherony, M. Khare, C. Lage, J. Pape, J. Sudijono, A. L. Steegen, and S. Stiffler. Scaling of 32nm low power SRAM with high-K metal gate. In *Proc. IEEE International Electron Devices Meeting IEDM 2008*, pages 1–4, 15–17 Dec. 2008.

[95] Kevin Zhang, U. Bhattacharya, Zhanping Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Yih Wang, Bo Zheng, and M. Bohr. A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply. *IEEE J. Solid-State Circuits*, 41(1):146–151, Jan. 2006.

[96] Song Zhao, Shaoping Tang, M. Nandakumar, D.B. Scott, S. Sridhar, A. Chatterjee, Youngmin Kim, Shyh-Horng Yang, Shi-Charng Ai, and S.P. Ashburn. GIDL simulation and optimization for 0.13 1.5 V low power CMOS transistor design. pages 43 – 46, 2002.

[97] Wei Zhao and Yu Cao. New generation of predictive technology model for sub-45nm design exploration. In *Proc. 7th International Symposium on Quality Electronic Design ISQED '06*, pages 6pp.–590, 27–29 March 2006.

[98] J. F. Ziegler. Terrestrial cosmic ray intensities. *IBM Journal of Research and Development*, 42(1):117 –140, jan. 1998.