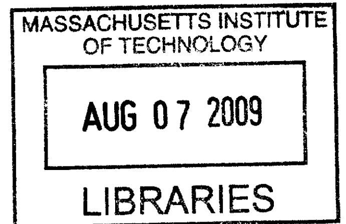
Ultra-Low-Power SRAM Design In High Variability Advanced CMOS



by

Naveen Verma

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

Doctor of Philosophy

ARCHIVES

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

© Massachusetts Institute of Technology 2009. All rights reserved.

Author ....
Department of Electrical Engineering and Computer Science
May 5, 2009

Certified by ....
Anantha P. Chandrakasan
Joseph F. and Nancy P. Keithley Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
Terry P. Orlando
Chairman, Department Committee on Graduate Theses

Ultra-Low-Power SRAM Design In High Variability Advanced CMOS

by

Naveen Verma

Submitted to the Department of Electrical Engineering and Computer Science
on May 5, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Embedded SRAMs are a critical component in modern digital systems, and their role is preferentially increasing. As a result, SRAMs strongly impact the overall power, performance, and area, and, in order to manage these severely constrained trade-offs, they must be specially designed for target applications. Highly energy-constrained systems (e.g. implantable biomedical devices, multimedia handsets, etc.) are an important class of applications driving ultra-low-power SRAMs.

This thesis analyzes the energy of an SRAM sub-array. Since supply- and threshold-voltage have a strong effect, targets for these are established in order to optimize energy. Despite the heavy emphasis on leakage-energy, analysis of a high-density 256×256 sub-array in 45nm LP CMOS points to two necessary optimizations: (1) aggressive supply-voltage reduction (in addition to V_t elevation), and (2) performance enhancement. Important SRAM metrics, including read/write/hold-margin and read-current, are also investigated to identify trade-offs of these optimizations.

Based on the need to lower supply-voltage, a 0.35V 256kb SRAM is demonstrated in 65nm LP CMOS. It uses an 8T bit-cell with peripheral circuit-assists to improve write-margin and bit-line leakage. Additionally, redundancy, to manage the increasing impact of variability in the periphery, is proposed to improve the area-offset trade-off of sense-amplifiers, demonstrating promise for highly advanced technology nodes. Based on the need to improve performance, which is limited by density constraints, a 64kb SRAM, using an offset-compensating sense-amplifier, is demonstrated in 45nm LP CMOS with high-density $0.25\mu m^2$ bit-cells. The sense-amplifier is regenerative, but non-strobed, overcoming timing uncertainties limiting performance, and it is single-ended, for compatibility with 8T cells. Compared to a conventional strobed sense-amplifier, it achieves 34% improvement in worst-case access-time and 4x improvement in the standard deviation of the access-time.

Thesis Supervisor: Anantha P. Chandrakasan
Title: Joseph F. and Nancy P. Keithley Professor of Electrical Engineering

Acknowledgments

MIT is truly a unique and wonderful place on this earth. For a new graduate student, as I once was, it can easily be too wonderful and too big. The only way to realize your place at MIT is through the guidance, encouragement, support, and friendship of an outstanding advisor like Prof. Anantha Chandrakasan. First and foremost, I thank Anantha. When I arrived here, I was not sure what, if anything, I could accomplish. Anantha, convinced me, by always expecting more from me, by always challenging me, and by supporting me through every research endeavor, that I could be a contributing member of this great community. His lessons for me have gone far beyond circuits; he has taught me to be a critical, sincere, cooperative, and respectful researcher. Anantha works firstly for his students, and I have learned more by watching him than I ever will from reading volumes of journals. As I proceed in my career, Anantha will always play an important role; he has given me something to strive for technically and personally. Thank you, Anantha, for your always strong support and guidance.

I am eternally grateful to my thesis committee members, Prof. Charlie Sodini and Prof. Duane Boning. Every researcher offers his work to the community hoping it is received by someone. To be able to discuss my work with such outstanding researchers as Charlie and Duane is the greatest honor of my career. Charlie and Duane have given this thesis a level of attention that has made the effort more than worthwhile. Thank you for your feedback and support, which has always aimed to make this thesis better. Because of your input, I am much prouder of this work, and after the many years it has consumed, that means a lot!

There are several faculty at MIT who have had a profound impact on me both technically and non-technically. I am extremely grateful to Prof. Harry Lee, who's mastery of circuits, and the ability to make that mastery accessible, has inspired me to study every last aspect of my field. I am grateful to Prof. Al Oppenheim who, by example, has shown me the impact that excellence in teaching can have and the level of dedication that must applied. I thank Prof. John Guttag for encouraging me to enthusiastically and intrepidly venture into new fields to seek out for myself how I

might broaden my contributions. Finally, I thank Prof. Joel Dawson for showing me that a newbie can have as big an impact as anyone, and he can do so without strain or tension, smiling all the way.

By far the most rewarding aspect of MIT has been the people I have been so fortunate to interact with. First, I must thank Margaret, who has repeatedly rescued me from overloads and crises. Margaret keeps ananthagroup running straight even when us students have accidentally gone in the wrong direction! Technically, the most fun I have ever had was discussing, debating, and pondering with Brian Ginsburg on matters of how to design an ADC (yes, many of the problems we hotly contested were already solved, but sometimes re-inventing the wheel is an unmatchable learning exercise!). I will always remember those years spent with Brian twisting my brain in front a white-board. Past members of ananthagroup, especially Benton Calhoun and David Wentzloff, showed me the ropes of being a graduate student. This, as they taught me, involves more than just tape-outs and paper deadlines; it involves lunch-time business plans, political/social debates, "useless" riddles and anecdotes, and most of all, laughs wherever they can be found. Also in this category are Alice Wang, Frank Honore, Fred Lee, and Raul Blazquez.

I am privileged to have the current members of ananthagroup around me every-day. I am especially grateful for the technical discussions and collaborations of Joyce Kwong, Yogesh Ramadass, and Nigel Drego (I will have more to say about these last two clowns shortly). I must thank my good friend Manish Bhardwaj, not just for his technical feedback but also for his support and encouragement, which was always on-hand when I needed it most (like when he put in a late night of chip testing with me to get results that were due the previous week!). Daniel Finchelstein, Denis Daly, and I arrived at MIT together, and I have had these two to lean on throughout my time here. They are the best fellow travellers one can hope for on this sort of journey, and I am grateful for their friendship the whole way through. It is also inspiring to see the newer students in the group, Vivienne Sze, Mahmut Ersin Sinangil, Patrick Mercier, and Masood Qazi, excelling and indeed becoming leaders.

I have especially been looking forward to say something about my friend Ali Shoeb.

His hyperactivity and enthusiasm are the main reasons why I will continually seek to expand and broaden my horizons beyond any narrow expertise I might have. Ali is genuinely inspired, and he inspires me! Eugene Shih is more controlled, but he has contributed equally to the fun I have had on the ninth flour of Building 32!

Thankfully, my experiences at MIT have actually gone far beyond MIT. I am extremely grateful for the support and encouragement I have received from collaborators at Texas Instruments. Most of all, Dennis Buss has been a champion of my work throughout my Ph.D. years. His enthusiasm has been a constant driving force, and he has spun miracles for me on more than one occasion to overcome the barriers and hurdles that inevitably arise during research. I am also grateful to Ted Houston, Wah-Kit Loh, Xiaowei Deng, Mike Clinton, Hugh Mair, and Alice Wang for their constant support and feedback.

I am thankful to Intel for providing me with fellowship support during my Ph.D. Even more importantly, Kevin Zhang of Intel has played a major role in how I have approached SRAMs from the research perspective. In fact, much of the work in this thesis has been inspired by his own research and the feedback he has been so generous to me with. Kevin has been a constant supporter and a mentor who I will always look to for stimulating discussions and input.

I am also thankful to Peter Holloway of National Semiconductor. It is much easier to do research when one has the kind of support that Peter has given me throughout my Ph.D. Peter has a unique perspective on circuits that is rooted in real-life; the only way a novice like myself can appreciate such a perspective is through the very intriguing and stimulating discussions I have had with him.

Completing a Ph.D. is far more than a test of technical execution. In fact, most of all, it is a test of will and morale. For both of these I am eternally grateful to the close friends I have made during my time here at MIT. Some of my most important moments at MIT have been spent during coffee-time with Nigel Drego and Yogesh Ramadass. Here, we got to transfer our analysis skill to all of life's great problems. None of us knows if we ever came close or even began to solve any of these, but we always returned from coffee less stressed, more motivated, and of course slightly

more awake... any way you cut it coffee-time is indispensible! Yogesh, Nigel, Vidya, Anand, and Nammi are great friends, and we are truly blessed to be able to laugh, lounge, and talk smack with them. The same, of course, goes for Daniel and Tarik (and Minou!). Since I arrived here at MIT Raj, Ferdi, Federico, and Gabi have been the rough-around-the-edges group with whom I could always be myself. This turns out to be a critical outlet when the pressure begins mounting, as it frequently does at MIT.

Finally, I come to my family, without whom nothing in my life, let alone my research, could ever have been possible. Most of all, my hard work and sincere efforts are for Mom Ji and Dad Ji. I have always relied on your love and prayers to lift me over obstacles. Of course, Vancouver is a continent away, but I have always felt you here with me, and that has been the strength I have needed. This thesis is for both of you. Thank you for your support, love, and blessings.

So far as effort put into this thesis is concerned, the first credit undoubtedly goes my amazing wife Anita. Ana, you are the reason behind this accomplishment, and your smile (and occasional craziness!) are the only rewards I hope for every day. Thank you for your love and support. I love you with all my heart.

I am blessed to also have the support and love of a second set of parents. Mom and Bug, thank you for your prayers, wishes, jokes, and love. I do not expect you to read this thesis, but I do hope you realize the role you have played in supporting me towards its completion. Thank you, once again, for your support, love, and blessings.

I am anxious to thank Angelee, Serena, and Jaimini. You three remind me that there is a lot more to my life than whatever I am busy with today. Thank you for the relief and lightening that your support and love always provides. This thesis truly could not have been completed without the formidable force behind me that you three have always been.

Similarly, Ang, Jason, and Connor, I know that you are always behind me and Ana, and we are externally grateful for the love, laughs, and lessons (about leather-backed turtles, etc.) that you have always provided.

Contents

1 Introduction **21**

1.1 Ultra-Low-Power Embedded SRAM Applications 24

1.2 SRAM Structure and Limitations 28

1.3 Thesis Contributions 31

2 SRAM Energy and Operating Metrics **35**

2.1 SRAM Energy 36

2.1.1 SRAM Idle-Mode Leakage Reduction 40

2.1.2 SRAM Sub-Array Optimal Energy 43

2.2 SRAM Operating Margins and Metrics 57

2.2.1 Read-Margin 58

2.2.2 Write-Margin 62

2.2.3 Hold-Margin (and Data-Retention-Voltage) 62

2.2.4 Cell Read-Current 65

2.3 SRAM Energy with Variation 65

2.4 Summary and Conclusions 70

3 Ultra-Low-Voltage SRAM Design **73**

3.1 Low-Voltage SRAM Challenges 75

3.1.1 Low-Voltage Bit-Cell Array 77

3.1.2 Low-Voltage Periphery 89

3.2 Ultra-Low-Voltage SRAM Prototype 91

3.2.1 8T Bit-Cell with Low-Voltage Circuit Assists 92

3.2.2 Sense-Amplifier Redundancy 100
3.2.3 Test-Chip Architecture 110
3.2.4 Measurements and Characterization 110
3.3 Summary and Conclusions 113

4 Performance Enhancement for High-Density SRAMs 117
4.1 High-Density SRAM Performance Challenges 118
4.1.1 Bit-Cell Read-Current 118
4.1.2 Sense-Amplifier Delay and Uncertainty 121
4.2 Single-Ended Sensing 124
4.3 High-Density SRAM Prototype 126
4.3.1 Non-Strobed Regenerative Sense-Amplifier 127
4.3.2 Test-Chip Architecture 145
4.3.3 Measurements and Characterization 148
4.4 Summary and Conclusions 150

5 Conclusions 153
5.1 Summary of Contributions 153
5.2 Concluding Thoughts and Future Directions 156

6 Appendix A: Acronyms 161

List of Figures

1-1 SRAM bit-cell density versus technology node showing cell density scaling in-line with transistor dimension scaling (every two years corresponds to a new technology node). 22

1-2 Three example low-power applications demonstrating dominating area and power-consumption of SRAMs: 45nm Intel Core 2 [9], 90nm ARM1176JZ (suitable for iPhone application processor) [10], and 65nm custom MSP430 [11]. 23

1-3 SRAM trade-offs. 24

1-4 Die photo of ultra-low-power low-voltage MSP430 microcontroller dominated by on-chip SRAM cache [11]. 27

1-5 Operating states of an SRAM where data-retention consumes energy even in the absence of active accesses. 29

1-6 Typical structure of modern SRAM; 6T bit-cell is composed of NMOS driver and access devices and PMOS load devices. 29

1-7 six-transistor SRAM bit-cell (6T) bit-cell butterfly curves showing bistable behavior during (a) hold, where access devices are "off", and during (b) read, where access devices are "on" and bit-lines are clamped to V_{DD}. 30

2-1 Simulated total leakage-current for 1Mb array in 45nm LP CMOS (at 1.1V); result shown includes variation and is normalized to total nominal leakage-current. 37

2-2 Active- and leakage-energy profiles in digital circuits showing trends expected in SRAMs. 39
2-3 Summary of parameters relevant to SRAM energy. 41
2-4 Normalized leakage-current reduction with respect to supply voltage for minimum-sized 90nm, 65nm, and 45nm devices due to DIBL (predictive models used). 42
2-5 Circuitry to enforce idle-mode biasing using (a) programmable sleep switches [63] and (b) an operational-amplifier [64]. 42
2-6 Waveforms corresponding to idle-to-active and active-to-idle mode transitions. 43
2-7 Summary of SRAM energy components. 44
2-8 Sub-array specifications for energy analysis. 49
2-9 Sub-array individual energy components. 52
2-10 Sub-array total energy (at room temperature) for various performance requirements (specified by $T_{CYC,RTN}$). 55
2-11 Energy components for $T_{CYC,RTN} = 10ms$ along $V_t = 0.45V$ axis. . . . 56
2-12 Mean and 4σ drain-current for minimum sized NMOS in 45nm CMOS with respect to (a) V_{DD} (with V_t=0.3V) and (b) V_t (with V_{DD}=1V). . 59
2-13 Read SNM definition through butterfly plots. 60
2-14 45nm $0.25\mu m^2$ bit-cell read SNM contours for (a) mean case, and (b) 4σ (on top of global variation) case. 61
2-15 45nm $0.25\mu m^2$ bit-cell write-margin contours for (a) mean case, and (b) 4σ (on top of global variation) case. 63
2-16 Hold SNM definition through butterfly plots. 64
2-17 45nm $0.25\mu m^2$ bit-cell hold SNM contours for (a) mean case, and (b) 4σ (on top of global variation) case. 66
2-18 45nm $0.25\mu m^2$ bit-cell read-current contours (log-magnitude) for (a) mean case (b) 4σ. 67
2-19 Sub-array total energy (at room temperature, with variation) for various performance requirements (specified by $T_{CYC,RTN}$). 69

3-1 Minimum supply-voltage of specifically ultra-low-voltage designs recently reported [84]. 74
3-2 Degradation of LP 65nm NMOS (predicitive model) with respect to V_{DD} showing (a) drain-current variation and (b) I_{ON}/I_{OFF}. 76
3-3 6T bit-cell for low-voltage analysis. 77
3-4 $0.5\mu m^2$ 6T bit-cell degradation of (a) read/hold SNM and (b) write-margin with respect to V_{DD}. 78
3-5 Electrical-β ratio definition and degradation with respect to V_{DD}. . . . 79
3-6 Bit-line leakage during read-data sensing opposing the ability to detect differential droops. 80
3-7 Read-current degradation in the presence of variation (a) with respect to V_{DD} scaling and (b) leading to loss of data sense-ability due to bit-line leakage. 82
3-8 Non-buffered bit-cells formed by (a) asymmetrically upsizing one pull-down path for rapid $RdBLT$ discharge [97], and (2) addition of device ($M7$) to gate bit-cell feedback path against disruption [98]. 84
3-9 8T bit-cell and layout (to overcome read-data-disruptions) shown besides a typical 6T bit-cell and layout. 86
3-10 6T bit-cell and 8T bit-cell operating margins for various size layouts (and equivalent read-current) in LP 65nm CMOS. 87
3-11 Bit-cell read-buffer enhancements to manage bit-line leakage using (a) PMOS/NMOS threshold-voltage skews [101], and (2) active pull-up on internal NCB node [102]. 88
3-12 8T bit-cell uses two-port topology to eliminate read SNM and peripheral assists, controlling $BffrFt$ and VV_{DD}, to manage bit-line leakage and write errors. 92
3-13 Read-buffer bit-line leakage in (a) conventional case where unaccessed read-buffer foot is statically connected to ground and (b) this design where unaccessed read-buffer foot is pulled up to V_{DD}. 93

3-14 $BffrFt$ driver must sink the read-current from all bit-cells in accessed row, and it draws leakage-current in all unaccessed rows. 94
3-15 To resolve read-buffer footer limitation (a) charge-pump circuit is used (b) BFB node gets bootstrapped to approximately $2V_{DD}$ increasing the current of the $BffrFt$ driver by over 500x. 95
3-16 Minimum word-line voltage resulting in a successful write with respect to the bit-cell supply voltage. 96
3-17 Virtual V_{DD} scheme (a) supporting circuits, and (b) simulation waveforms. 97
3-18 Read-current gain as a result of read-buffer upsizing (a) via width increase, and (b) via length increase (taking advantage of reduced variability and RSCE). 99
3-19 8T bit-cell layout with read-buffer upsizing and $BffrFt$ control (but no VV_{DD} control). 100
3-20 Final 8T bit-cell layout and folded-row tiling. 100
3-21 Differential sense-amp structure cancels effects of global variation. . . 102
3-22 Monte Carlo simulations of sense-amp statistical offset; at expected input swing (i.e. 60mV), errors from offset are prominent. 103
3-23 With sense-amplifier redundancy, each $RdBL$ is connected to N different sense-amplifiers. 103
3-24 With sense-amplifier redundancy (a) the size of each individual sense-amplifier must decrease, and (b) the individual sense-amplifier error probabilities, defined as the area under the offset distribution exceeding the magnitude of the input swing, increases. 104
3-25 Increased levels of redundancy significanly reduce the error probability in the overall sensing network. 105
3-26 Redundancy selection circuitry consisting of a dummy bit-cell and selection state-machine. 106
3-27 Overall error probability for implemented sense-amp redundancy scheme improves by a factor of 5 compared to a single sense-amp scheme. . . 107

3-28 Sense-amplifier redundancy overhead circuitry for the case of $N = 2$. . 108
3-29 Normalized sensing-network ($N = 2$) error probabilities for different technologies and layout areas. 109
3-30 Prototype test-chip architecture, with total capacity of 256kb partitioned in eight sub-arrays. 110
3-31 Die photo of prototype low-voltage SRAM. 111
3-32 Prototype SRAM leakage-power; at the minimum V_{DD} of 0.35V, the entire SRAM draws 2.2μW of leakage-power. 112
3-33 SRAM speed with respect to V_{DD}. 113
3-34 Total power (solid curves) and leakage power (dotted curves) with respect to operating frequency. 113

4-1 Degradation in bit-line discharge time for high-density SRAMs caused by (a) reduced cell read-current and (b) increased bit-line capacitance. 119
4-2 Read SNM trade-off in high-density SRAMs limited by (a) cell size and (b) inverse correlation with cell read-current, caused by opposing access-device requirements. 120
4-3 Conventional strobed sense-amplifier topologies with (a) one input-output port and (b) separate input-output ports. 121
4-4 Array read-path and sense-amplifier strobe-path (a) limited by matching to 5σ bit-cell and (b) exhibiting severe delay divergence over process-voltage-temperature conditions, leading to excess overall delay. 125
4-5 Non-strobed regenerative sense-amplifier (NSR-SA) schematic and ideal transfer function. 127
4-6 NSR-SA circuit and waveforms during reset phase. 129
4-7 NSR-SA circuit and waveforms during detection phase (for both bit-line logic cases). 130
4-8 Output clocking (a) at array-level with (b) waveforms showing decoupling from internal critical read-path. 131
4-9 Offset compensation (a) technique and (b) analysis. 132

4-10 10k point Monte Carlo simulation showing improved sigma of NSR-SA access-time compared to conventional sense-amplifier access-time. . . 134

4-11 NSR-SA robustness to false-regeneration in the presence of charge-injection errors. 136

4-12 NSR-SA technique to set regeneration trip-point (V_{TRIP}) for noise-rejection and sensitivity considerations. 137

4-13 NSR-SA (a) circuit showning noise sensitive nodes (X/Y), and (b) reponse of X/Y due to transient spikes on V_{DD}, and (c) Response of X/Y leading to output errors on QB due to sustained step on V_{DD}. 138

4-14 NSR-SA noise measurement simulation setup. 139

4-15 Example bit-line noise sources originating (a) from precharge, word-line, and column-select control signal coupling, and (b) substrate coupling. 140

4-16 NSR-SA input transfer characteristic. 142

4-17 Input transfer characteristic for (a) inverter and (b) two stage inverter cascade. 142

4-18 NSR-SA V_{DD} noise transfer characteristic. 143

4-19 NSR-SA input transfer characteristic with ±50mV V_{DD} noise. 143

4-20 NSR-SA transfer characteristic for (a) V_{SS} noise and (b) input with ±50mV V_{SS} noise. 144

4-21 Input errors resulting from V_{DD} and V_{SS} noise. 145

4-22 Block-diagram of prototype test-chip and access-time measurement methodology. 146

4-23 Dedicated circuitry to inject a controllable noise-amplitude on one set of bit-lines and independently adjust the sensitivity/noise-rejection of the NSR-SA. 147

4-24 IC die photo of prototype implemented in low-power 45nm CMOS to compare performance of NSR-SA with conventional sense-amplifier. . 148

4-25 Access-time measurements from 53 chips (at 1V) showing a factor of four improvement in the NSR-SA distribution sigma compared to the conventional sense-amplifier sigma. 149
4-26 Measured bit-line noise-rejection with respect to access-time, showing ability to tune one at the cost of the other. 149

List of Tables

1.1 Key existing and emerging applications for biomedical devices 25

1.2 Energy collecting and harvesting options [29][30][31][32] 27

4.1 Test-chip performance summary. 150

Chapter 1

Introduction

Moore's law of scaling [1] has been the most important driving force behind the semiconductor industry. Scaling has directly or indirectly been the root cause of the tremendous capabilities of today's ICs and their ubiquitous use in nearly all modern electronic systems. Though Gordon Moore recently amended his law to include a much broader set of metrics associated with ICs [2], his basic statement pertains to "components," which literally implies number of transistors. Today, even as many aspects of CMOS device scaling begin to saturate off the exponential trend, density-scaling remains a primary objective of the semiconductor industry [3]. In the face of rapidly emerging limitations that are fundamental to continued device shrinking, density-scaling enables circuit [4] and architecture level parallelism [5], providing a means to achieve energy-efficiency and performance improvements in lieu of of the previous trends.

Embedded SRAMs provide a direct means of bringing the benefits of transistor-level density-scaling to the circuit and architecture levels and are therefore vital to this new model of IC scaling. Due to their regular structure and broad applicability to so many digital systems, SRAMs are carefully designed as one of the lead components during the development of new technology nodes, and they utilize highly specialized and aggressive layout rules that address sub-resolution fabrication limitations. This level of design attention has allowed SRAM bit-cells to follow density trends in-line with the transistors themselves [6]. This is shown in Figure 1-1 where bit-cell areas

reported by Intel, IBM, TI, Sony, Renesas, and Samsung have been plotted versus the technology node (represented by deployment year).

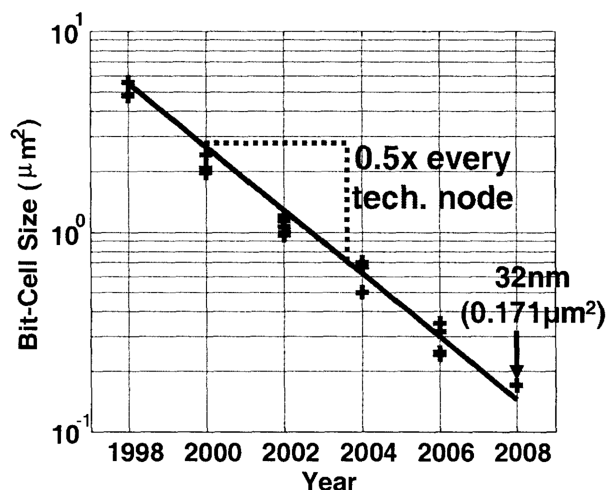


Figure 1-1: SRAM bit-cell density versus technology node showing cell density scaling in-line with transistor dimension scaling (every two years corresponds to a new technology node).

Accordingly, to benefit efficiently from transistor density-scaling, modern digital architectures increasingly emphasize the use and integration of more and more SRAMs [7][8]. The resulting consequence for low-power devices is that SRAMs occupy a dominating portion of the total die area and the total power consumption. Figure 1-2 shows three state-of-the-art examples intended for increasingly low-power applications: the Intel Core 2 processor targets mobile computing [9], the ARM1176JZ processor targets hand-held computing [10], and the custom MSP430 microcontroller targets remote wireless sensor and implantable biomedical computing [11]. The important trend observed here is that the SRAM (or memory) power becomes more and more significant in increasingly low-power devices. The precise cause of this is discussed throughout the following chapters, but in the meantime, it is clear that SRAMs are a fundamental platform component in the modern semiconductor industry, and their power-consumption is a limiting factor.

An important evolution in the semiconductor industry is that, today, the application space for integrated circuits is extremely broad, extending far beyond desktop computing microprocessors to include ambient, remote, mobile, and implantable de-

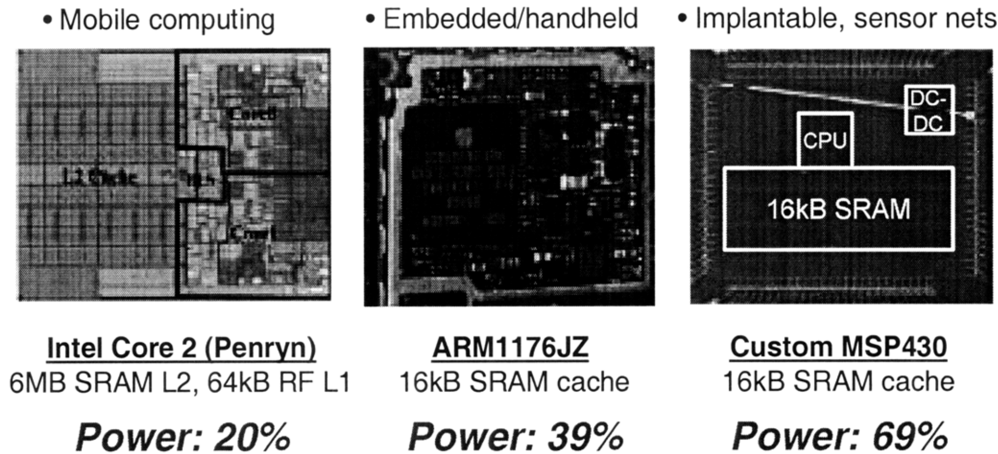


Figure 1-2: Three example low-power applications demonstrating dominating area and power-consumption of SRAMs: 45nm Intel Core 2 [9], 90nm ARM1176JZ (suitable for iPhone application processor) [10], and 65nm custom MSP430 [11].

vices, to name a few. With regards to the constituent digital circuits, all of these applications have vastly varying and highly stringent demands that require careful design within the associated trade-offs. In order to adhere to intense scaling trends, SRAM design is also highly constrained, especially in the face of emerging limitations ranging from device-level variability to system-level power consumption. Since their impact on the overall system is so significant, and since their design is so constrained, modern embedded SRAMs must be developed with the application in mind so that their own trade-offs can be carefully managed. Generally speaking, SRAMs are strongly subject to the power, performance, and density trade-offs shown in Figure 1-3. The precise origins and effects of these trade-offs are discussed throughout the following chapters, but the overall implication is that improvement in one of the dimensions strongly stresses the others. Of course, all three dimensions are important to some degree in all applications; as a result, embedded SRAM design involves making judicious compromises in order to support the most important system-specific requirements. The focus of this work is to investigate techniques that improve the basic trade-off in order to more efficiently allow optimization of the parameters relevant for the systems considered (these are discussed in more detail below). It is important to note that although the illustration in Figure 1-3 indicates a simple inverse relation-

ship between power, performance, and density, in reality, the relationships are often much more complicated, and, importantly, aggressive emphasis on one dimension, such as power reduction, increases the opposition imposed by the other dimensions with much higher intensity.

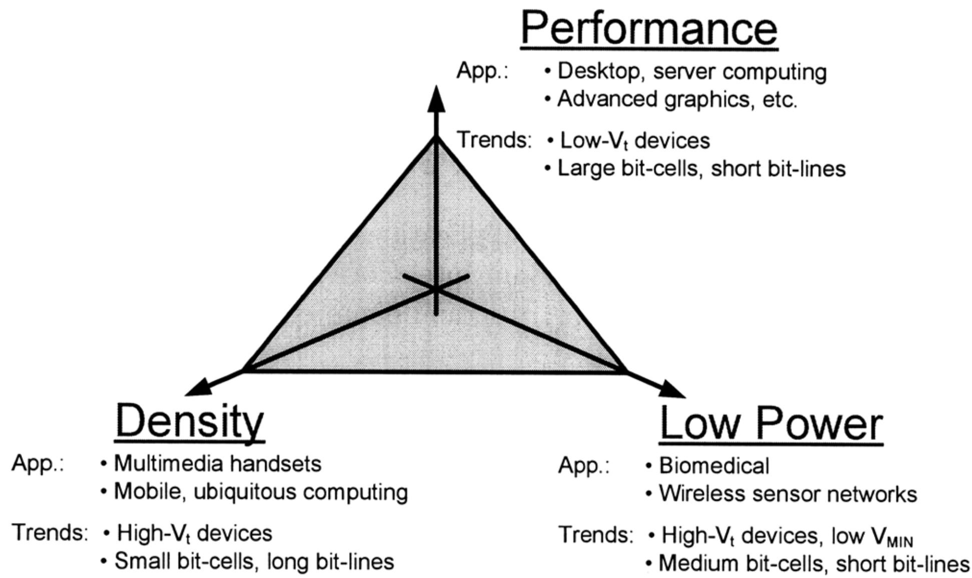


Figure 1-3: SRAM trade-offs.

1.1 Ultra-Low-Power Embedded SRAM Applications

Since SRAMs must be specially designed with their application in mind, it is worth considering the application constraints. This work specifically considers a number of applications where power consumption, or, more generally, energy consumption, is paramount. Of course, the SRAM challenges associated with achieving multi-Giga-Hertz operation in high-performance applications, including desktop and server computing, requires very targeted and innovative solutions as well [12][13][14]. However, a few of the highly energy-constrained applications that are the focus of this work are considered below:

Table 1.1: Key existing and emerging applications for biomedical devices

Application	**Performance Specification**		
	Power	**Processor**	**Energy Source**
Pacemaker & Cardioverter-defibrillator [15][16]	<10 μW	1kHz DSP	10-year lifetime battery
Hearing aid & Cochlear implant [17][18][19]	100-2000 μW	32kHz-1MHz DSP	1-week lifetime battery
Neural recording [20][21]	1-10 mW	n/a	Inductive power
Body-area monitoring [22]	140 μW	<10MHz DSP	Battery

(1) **Biomedical.** Existing and emerging biomedical applications are shown in Table 1.1, along with some critical system requirements. In all cases, energy is highly constrained. In the case of implantable devices, such as pacemakers/defibrillators, cochlear implants, and neural sensors/stimulators, battery lifetime constraints determine the time between surgical replacement, thereby limiting total system power consumption to $100\mu W$ or less. Wearable systems, such as hearing aids and body-area sensors, have similar, though somewhat less stringent, energy-constraints set by battery weight limitations.

Although the energy constraints in biomedical systems are severe, their performance constraints are considerably relaxed. Table 1.1 shows that, for the most part, processors need to operate at less than 1MHz. Additionally, the volume of most of the highlighted applications is fairly modest, though it does range to much higher volumes as well, especially in the case of body-area sensors. Correspondingly, the required SRAMs must heavily emphasize low-power and, secondarily, density, and these can be optimized at the cost of performance.

(2) **Mobile multimedia.** Today's portable handsets are capable of extremely sophisticated multimedia. In addition to rich audio and communication capabilities, they will deliver high-definition video to users [23]. For these applications,

however, the time required between battery charges must be extended to the order of several days, and the battery itself can weigh no more than a few tens of grams. Accordingly, power consumption is a major concern, though it is not as constrained as in biomedical systems. Also unlike biomedical applications, performance is critical in order to support rich multimedia operations, that require processors with operating frequencies up to hundreds of Mega-Hertz. Further, the large volume of consumer handsets implies that cost and density are also primary concerns. As a result, very high-density SRAMs that minimize power consumption under moderate-to-high performance constraints are needed.

(3) **Wireless sensor networks.** Micro/nano-scale devices providing sensing, processing, and communications capabilities can form networks, broadly referred to as wireless sensor networks [24][25]. The applications for such devices include industrial and automotive sensing [26], environment monitoring [27], structural monitoring [28], and military surveillance/detection. Operation of such networks must be largely maintenance-free due to their use in remote or inaccessible physical locations. As a result, battery lifetime constraints are critical, and the battery must be physically small to facilitate in-situ sensing in a broad range of uses. Alternatively, to extend the lifetime of the sensor nodes, potentially indefinitely, energy harvesting from the ambient environment can be leveraged as long as occasional degradation in performance quality, depending on the ambient factors, can be tolerated. Nonetheless, the power consumption of the system is limited by the harvesting capacity. Table 1.2 shows the power harvestable by state-of-the-art energy harvesting devices, indicating a total power budget less than $100\mu W$ for most of the sensor networks considered.

With regards to SRAM requirements, power consumption (both static and dynamic) is the primary concern, and, since most monitoring applications require processing on low-speed signals, performance constraints are relaxed to the hundreds of kilo-Hertz range. Since the nodes are meant to form high-density networks that are sacrificial after use, cost and density are also important concerns.

Table 1.2: Energy collecting and harvesting options [29][30][31][32]

Energy Source	**Performance**
Thermoelectric	60 μW/cm^3
Light	100 μW/cm^2 (office), 100 mW/cm^2 (direct light)
Vibration	4 μW/cm^3 (human motion)
Heel strike	10-700 mW (walking)
Near-field inductive energy transfer	20 mW at 5 cm [33]
Far-field inductive energy transfer	2 μW at 10 m [34]

As with most digital systems, embedded SRAMs play a highly prominent role in these energy-constrained applications. Also as before, they pose the most critical limitation to the total power, performance, and area. Figure 1-4 shows an example of a custom MSP430 microcontroller that specifically targets highly energy-constrained biomedical and sensor applications [11]. Operating at its minimum energy point, its on-chip SRAM cache consumes 69% of the total energy per operation, limits the operating frequency (which is 1.7MHz for the SRAM at 0.5V), and, as shown, occupies a dominating portion of the total area. Consequently, to enable the applications described above, embedded SRAM is a critical area of focus.

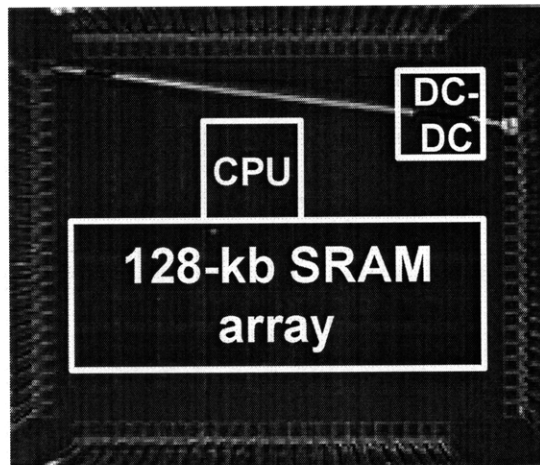


Figure 1-4: Die photo of ultra-low-power low-voltage MSP430 microcontroller dominated by on-chip SRAM cache [11].

Energy Versus Power

For the applications discussed above, it is important to make the distinction between energy consumption and power consumption. Ultimately, battery powered systems are primarily limited by the energy the battery can provide. Energy harvesting systems typically use a battery (or other form of energy storage [35]) to buffer the power extracted from an ambient source [36], and, once again, average power consumption, corresponding to total energy normalized over a time period, is the critical concern. Performing any circuit operation requires energy, and, so, it is a fundamental metric for battery operated and energy-harvesting systems.

This implies that in an "off" state, where the circuit is performing *no* operation, it can consume extremely low energy. Such an "off" state, however, only exists in very specific cases for SRAMs. Generally, even in the absence of active accesses, SRAMs are expected to retain their stored data. Figure 1-5 shows this distinction, and, in the case of the persistent storage states, data retention is an operation that requires energy. Importantly, however, this operation is inherently tied to time by the duration for which data retention is required. Of course, ultimately, the SRAM will transition to the "off" state, either at the end of the device's lifetime or upon completion of a set sequence of operations. Accordingly, the total energy can still be considered. However, unlike with generic digital logic, the energy consumed has a component related to time, but unrelated to the time associated with its own circuit delay. The corresponding energy optimization is considered in detail in Chapter 2.

1.2 SRAM Structure and Limitations

Figure 1-6 shows the architecture used by modern SRAMs. A combination of row decoders and column multiplexers provide access to the bit-cells. While data-retention circuits for logic, like flip-flops and latches, typically employ between 10 to 20 devices, the 6T bit-cell shown relies on ratioed operation to achieve the required functionality with very high density. 6T CMOS bit-cells in the 65nm and 45nm nodes occupy 0.4-0.5μm^2 [37][38] and 0.24-0.33μm^2 [39], respectively. For reasons explained below,

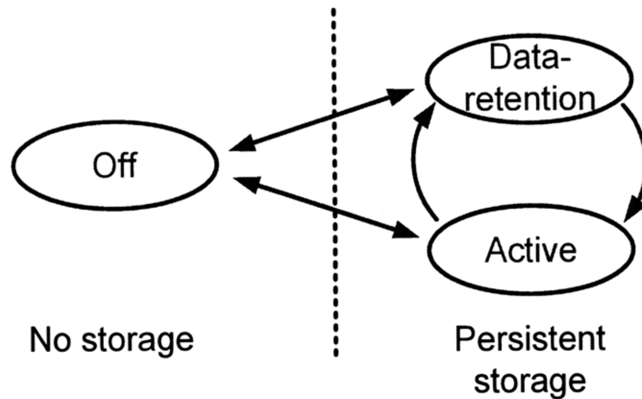
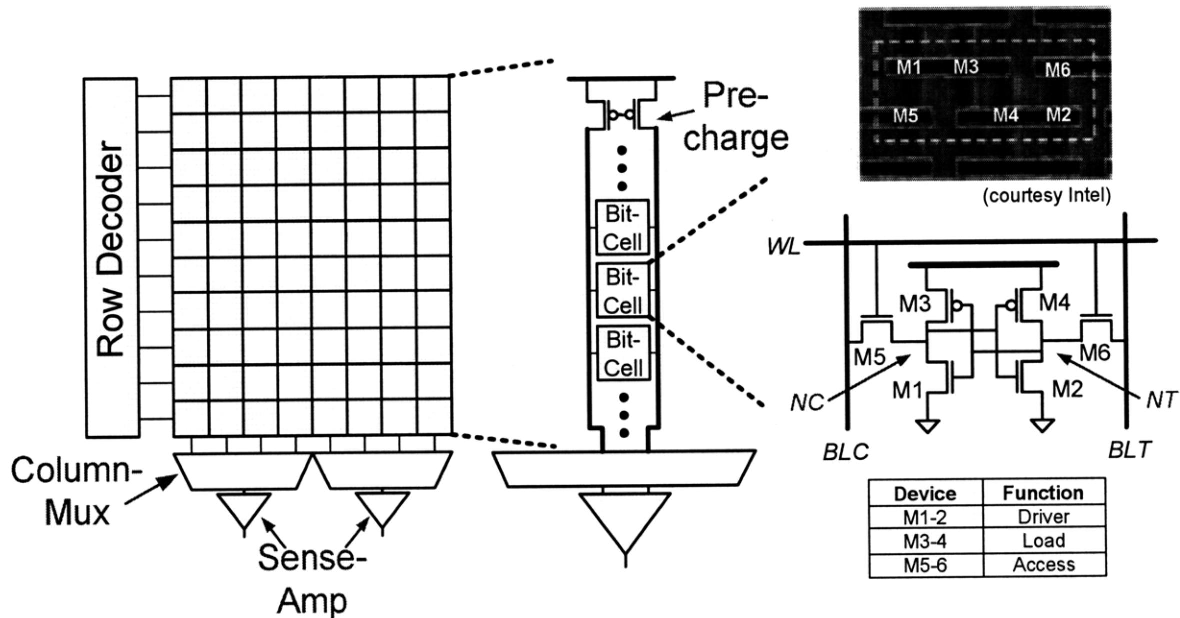


Figure 1-5: Operating states of an SRAM where data-retention consumes energy even in the absence of active accesses.

$M1-2$ are called the driver devices, $M3-4$ are called the load devices, and $M5-6$ are called the access devices.



Device	Function
M1-2	Driver
M3-4	Load
M5-6	Access

Figure 1-6: Typical structure of modern SRAM; 6T bit-cell is composed of NMOS driver and access devices and PMOS load devices.

Data is held in the 6T cell by the cross-coupled inverter structure (formed by $M1-4$). Figure 1-7a shows how the 6T cell's ability to hold data depends on its butterfly curves. Here, the transfer-functions between the data storage nodes, NT/NC, are superimposed, and the bi-stable nature required is indicated by intersection points at

valid logic "0" and "1" levels. Strictly speaking, read-access is a non-ratioed operation where the bit-lines, BLT/BLC, are precharged, and, after word-line (WL) assertion, the cell read current, I_{RD}, which is generated by the driver and access devices, causes a droop on one bit-line which can be sensed with respect to the other to quickly decipher the accessed data. However, the transients on NT and NC can result in loss of the bi-stable characteristic, and their worst-case impact can be analyzed by assuming that BLT/BLC are clamped at V_{DD}. The corresponding butterfly curves, shown in Figure 1-7b, now have dangerously degraded lobes, quantified by the static noise margin (SNM), which measures the diagonal length of the largest embedded square [40]. An SNM less than zero implies the loss of one of the required intersection points, indicating the cell's inability to correctly retain the corresponding data state. Hence, proper operation requires maintaining wide lobes, which depends on the driver devices, $M1-2$, being much stronger than the access devices, $M5-6$.

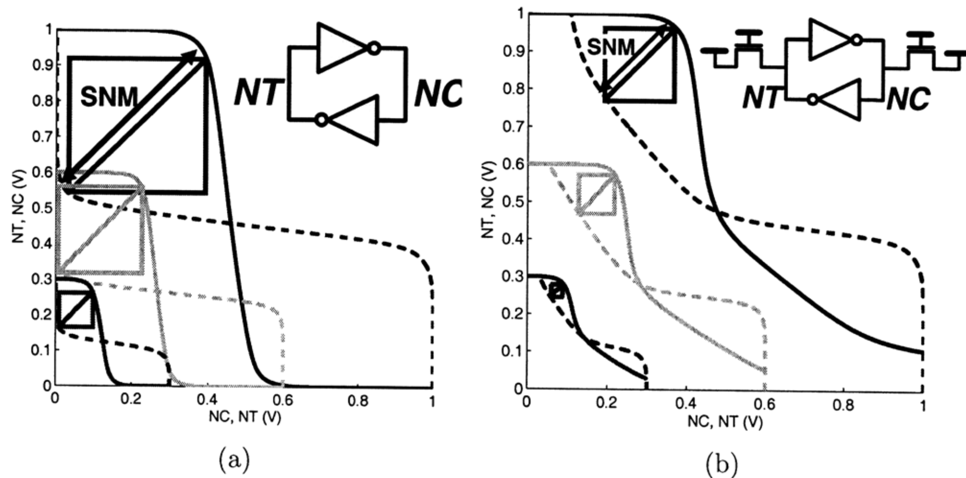


Figure 1-7: 6T bit-cell butterfly curves showing bi-stable behavior during (a) hold, where access devices are "off", and during (b) read, where access devices are "on" and bit-lines are clamped to V_{DD}.

Data is written to the 6T cell by pulling the appropriate bit-line low. The cell is made mono-stable at only the desired data value, and, after WL gets de-asserted, the local feedback regenerates to the correct state. Write operation is explicitly ratioed, since the NMOS access devices are required to overpower the PMOS load devices, $M3-4$, in order to overwrite new data.

SRAM Variation

The ratioed operation, both during read and write, leaves the 6T bit-cell highly susceptible to both variation and manufacturing defects. In particular, since a typical SRAM is composed of bit-cell arrays of hundreds of kilo-bits to several Mega-bits, extreme worst-case case behavior at the 4 or 5σ level must be considered.

Two forms of variation affect SRAMs: inter-die (which will be called global variation) and intra-die (which will be called local variation) [41]. Global variation is the difference between average parameter values of the die; for instance, these can include the average NMOS/PMOS threshold voltage, dielectric thickness, or poly width. Global variation comes about due to systematic processing changes affecting individual dies. On the other hand, local variation is the difference between nominally matched devices on the same die. These can include the number of NMOS/PMOS channel-adjust doping ions, poly line-edge roughness, local-layout-dependant lithography effects, as well as transient effects such as negative bias temperature instability (NBTI) [42]. In advanced technologies, local variation sources have an increasingly dominating impact [41]; while global variation significantly degrades the operating margins of SRAMs, local variation represents the most urgent concern regarding the increasing rate of failures observed [43]. A complete treatment of variation in CMOS devices, and its impact on circuits, such as SRAMs, can be found in [41].

1.3 Thesis Contributions

Previous work in SRAMs has focused on their reliability with technology and density scaling. The use and implications of technology optimizations that are generally pursued for a broad range of high-volume and low-energy applications (e.g. mobile processors) have also begun to be investigated. There remains, however, the need to develop SRAM techniques to support severely energy constrained applications such as biomedical devices, wireless sensor nodes, and much richer mobile multimedia. Specifically, these require strategies to improve the trade-offs highlighted in Figure

1-3.

Due to its heightening importance in digital systems, and its increasing sensitivity to processing and manufacturing factors, SRAM design requires some level of coordination with technology development in order to be effective. As a result, low-energy SRAM solutions must be compatible with industry methodologies, which are well suited for new technology development at the manufacturing level. For instance, optimal bit-cell layout design depends on several manufacturing details. Accordingly, this work focuses on circuit techniques that are compatible with and supportive of those approaches, particularly with regards to the most advanced technologies. It is the hope that this thesis contributes to identifying and solving some of the most critical issues facing highly energy constrained SRAMs, though, of course, many issues will remain, and every effort is made to identify those as well.

This thesis contributes in the following areas:

(1) **SRAM Energy Analysis.** Supply- and threshold-voltage strongly impact the total energy of an SRAM sub-array. Chapter 2 presents an analysis for the optimal supply-voltage (V_{DD}) and threshold-voltage (V_t) targets in order to minimize total energy considering the need to perform a given average number of accesses within a specified time. The analysis here is different from that of generic logic [44] in two ways: (1) the presumed need to retain the stored data for the entire time specified, and (2) the increased dependence of the energy on variation, which in SRAMs occurs at extreme-levels.

In addition to optimal targets from the perspective of minimizing energy, Chapter 2 considers how the metrics that are critical to SRAM operation depend on the supply- and threshold-voltage targets. As a result, the major oppositions to SRAM operation at the optimal energy point are established.

(2) **Ultra-Low-Voltage SRAM.** The analysis of Chapter 2 points to ultra-low-voltage operation as a means to minimize sub-array energy. Chapter 3 provides an analysis of failure sources within the SRAM that restrict low-voltage operation. Having analyzed the failure sources, techniques are proposed to overcome

them, and the techniques are analyzed for their efficiency. The techniques address two key limitations: (1) bit-cell operation and (2) sense-amplifier operation. Redundancy, which is commonly relied on to overcome bit-cell variation at the 5σ level, is analyzed for critical periphery components (namely, sense-amplifiers), where low-voltage operation exacerbates variation to an intolerable point even at the 3σ level. The proposed techniques are demonstrated in a prototype 256kb SRAM test-chip in 65nm LP CMOS that operates down to 0.35V.

(3) **Low-Power High-Density SRAM Performance Enhancement.** The analysis of Chapter 2 points to sub-array performance as a major limitation to energy reduction, especially in the presence of variation. Chapter 4 analyzes the severe trade-off between sub-array performance and density. The limitations imposed by both the bit-cells and the sense-amplifiers are investigated to alleviate the constraining trade-offs. Specifically, a sense-amplifier is proposed that provides regenerative small-signal sensing. Importantly, however, it does not require an explicit strobe signal, which, in advanced technologies, imposes severe timing uncertainties that limit the worst-case performace. Additionally, due to the promise of single-ended bit-cells (e.g. 8T) for ultra-low-voltage, low-energy applications, the sense-amplifier proposed provides variation resilient single-ended sensing. Although this enables the low-energy benefits of voltage scalability and high read-current, it introduces increased sensitivity to noise sources. Accordingly, the noise performance of the proposed sense-amplifier is analyzed. A prototype test-chip in 45nm LP CMOS compares its performance to that of a conventional strobed sense-amplifier, demonstrating improvements in the worst-case access-time and the standard-deviation of the access-time by 34% and 4x, respectively.

Chapter 2

SRAM Energy and Operating Metrics

With respect to the growing number of applications considered in Chapter 1 and the increasing dominance of SRAMs, careful consideration is required of the trade-offs that minimize SRAM energy. The aggressive application of these energy-reducing trade-offs, however, directly impacts the functionality and operating metrics of the SRAM (and, in turn, the system) leading to a complex effect on the achievable energy savings in a practical scenario. Of course, device variation, at the extreme levels observed in typical SRAM arrays, plays a central role in precisely how the energy-reducing trade-offs affect the operating metrics. Since their energy is so critical in the overall system, SRAMs are subject to a sophisticated suite of power-management assists spanning the device, circuit, and architecture levels. The energy, then, must be analyzed under this power-management strategy.

Both active- and leakage-energy components contribute critically to SRAM energy, and hence the analysis in this chapter treats them as the underlying optimization targets. For general digital circuits, it has already been shown that supply-voltage (V_{DD}) and threshold-voltage (V_t) interact to set the active and leakage energy [44]. Compared with general digital-circuits, however, SRAMs face the operational constraint of long-term data-retention even during temporary idle periods (that may last arbitrarily long) where it is known that active accesses will not be performed. This

gives rise to the concept of a data-retention voltage (V_{DRV}) [45], where only idle data-storage, and no data-read or data-write, functionality must be supported. In addition to their effect on energy, which is the primary motivation for manipulating V_{DD}, V_t, and V_{DRV}, this chapter analyzes the fundamental effect these voltages have on SRAM functionality and performance in the presence of variation. Ultimately, this chapter serves to determine what the optimal operating point (i.e. V_{DD} and V_t) target is to minimize SRAM energy and also to identify the challenges of operating at that point.

2.1 SRAM Energy

The array nature of SRAMs has an important impact on the way their energy scales with respect to V_{DD} and V_t, especially during active-access modes. Specifically, compared to general digital circuits, SRAM leakage-energy has increased importance due to three factors: (1) high ratio of leakage-paths to actively-switching-nodes, (2) total leakage set by an aggregation of intentionally minimum sized devices, and (3) critical-path set by a single MOSFET pull-down stack with extreme variation. These factors are considered below.

In order to maximize array area-efficiency, the trend is to use large sub-arrays with up to 256 bit-cells (or more) per row and column [46], as far as performance optimizations allow [47][48]. For such large sub-arrays, the leakage from the bit-cells, which scales directly with the array size, dominates over that of the periphery. Within the sub-array, the active switch capacitance from the word-lines scales with the number of columns but not the number of rows, since only one row's word-line switches per access. As a result, the word-line switch capacitance does not increase in proportion to the total array size. Alternatively, the switch capacitance of the bit-lines scales with the number of rows, and, during read-accesses, the bit-lines of all columns switch; however, typically, their swing is significantly less than V_{DD}. Further, during write-accesses, the number of bit-lines that switch is reduced by the column-multiplexer ratio (typically four or eight). Consequently, for large sub-arrays, the

ratio of leakage-energy to active-energy is higher than that of generic logic.

The use of intentionally small devices, to maximize the density of the bit-cell arrays, introduces increased variation, elevating the actual aggregate leakage-current significantly beyond the nominal aggregate leakage-current. Since leakage-current is related exponentially with threshold-voltage, the effect of V_t variation cannot be expected to average out over the linear summation of all leakage-paths in the array. Figure 2-1 shows the simulated total aggregate leakage-current (at 1.1V), normalized to the nominal aggregate leakage-current, for a 1Mb array composed $0.25\mu m^2$ bit-cells in an LP 45nm technology. As shown, increasing σV_t (even over a fairly modest range) leads to a significant increase in the total leakage-current [49]. To simplify the description, this will be referred to as the leakage-current gain factor due to variation.

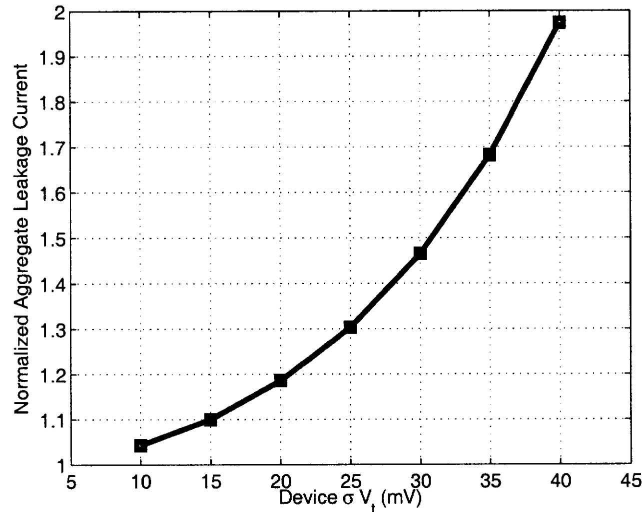


Figure 2-1: Simulated total leakage-current for 1Mb array in 45nm LP CMOS (at 1.1V); result shown includes variation and is normalized to total nominal leakage-current.

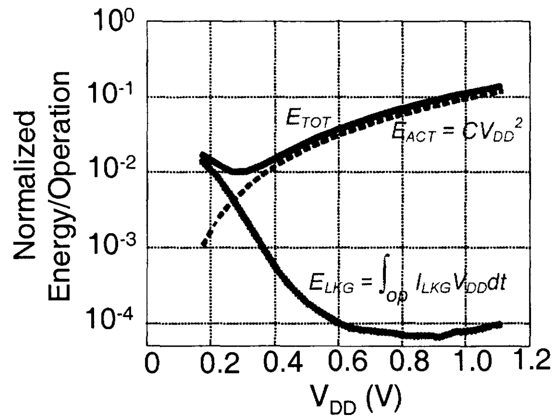
The critical delay path in an SRAM is limited by the time required for the accessed bit-cells to discharge their bit-lines beyond the required data-sensing margin. In the presence of variation, this implies that the performance of a large array may be set by a single bit-cell experiencing drive-current degradation at an extreme level (e.g. 5σ). The performance degrading effect of variation in a typical circuit composed of logic paths is alleviated since the total delays are set by the sum of several constituent

stages [50]. Consequently, extreme variation on any one device has greatly reduced impact. Unfortunately, in SRAMs the tendency towards large arrays implies the possibility of extreme variation, and the structure of the read-path precludes the benefit of delay averaging over many stages. As a result, the overall performance of an SRAM suffers far more drastically in the presence of variation.

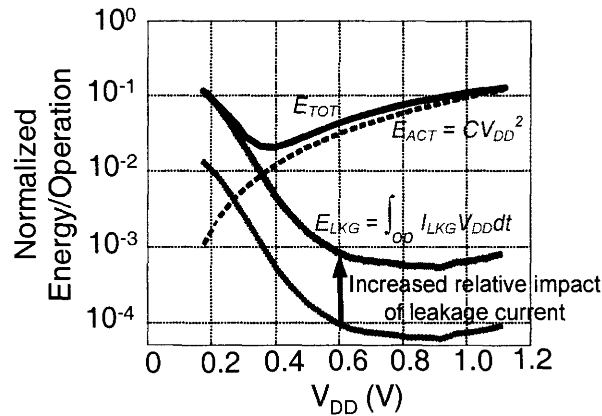
Considering the active and leakage energy profiles for a general digital circuit [51], the active-energy scales quadratically in a straight-forward manner as CV_{DD}^2 with respect to supply-voltage. Of course, as a circuit's V_{DD} is reduced, however, the gate-drive of the constituent MOSFETs is also reduced, degrading the switching speed. Consequently, the integration time of the leakage-currents, which is set by the time required to complete the operation, increases, raising the leakage-energy. The opposing active and leakage energy profiles are shown in Figure 2-2a for a representative case (i.e. 32b carry-look-ahead adder in 90nm CMOS).

However, based on the factors discussed above leakage-energy in SRAMs has increased prominence. Specifically, as sketched pictorially in Figure 2-2b, the high ratio of leakage-paths to actively-switching-nodes and the leakage-current gain factor due to variation both contribute to raising the leakage-energy curve up-ward relative to the active-energy curve. Additionally, the severe performance degradation due to the critical-path's dependence on a single bit-cell experiencing extreme variation, causes the leakage-energy curve to shift right-ward, as sketched in Figure 2-2c. This can be understood by observing that the point at which the leakage-energy begins increasing exponentially occurs at a higher supply-voltage than before; effectively, variation raises the limiting bit-cell's threshold voltage, and, as a result, supply-voltage reduction quickly leads to sub-threshold operation, which imposes an exponential increase in circuit delay.

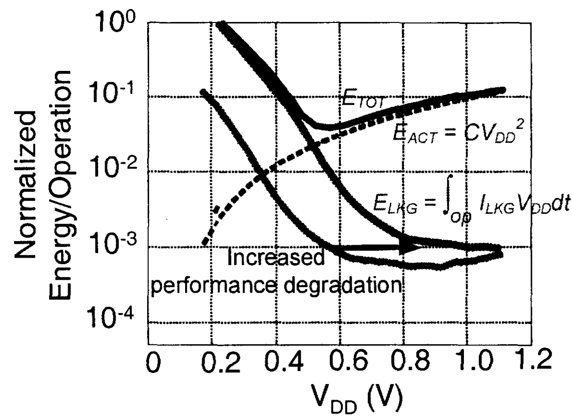
The result in Figure 2-2c seems to indicate that the optimal V_{DD} for SRAMs occurs at a relatively high supply-voltage. In fact, however, the energy optimization picture must be modified by considering the practical power-management approach discussed in Section 2.1.1. Although the importance of leakage-energy remains high, it must be considered both during active-access and idle-data-storage modes. As



(a) Energy profiles represntative of generic logic (90nm 32b carry-lookahead adder).



(b) Relative leakage-energy shift exepcted in SRAMs due to increased ratio of leakage-currents to active-switching-current.



(c) Relative leakage-energy shift expected in SRAMs due to severe performace degradation from bit-cell variation.

Figure 2-2: Active- and leakage-energy profiles in digital circuits showing trends expected in SRAMs.

discussed below, raising V_{DD} in order to reduce the SRAM access delay has reduced benefit, as leakage-energy must still be incurred in order to retain data even after the active-mode.

The following subsections start by describing the operating modes of an SRAM. Then, the energy components during these modes are identified and analyzed in detail, especially with respect to the supply- and threshold-voltages. Finally,V_{DD} and V_t targets are determined to optimize energy.

2.1.1 SRAM Idle-Mode Leakage Reduction

If the SRAM power-supply could be gated after the completion of a required number of accesses, the picture in Figure 2-2, consisting of one leakage energy component and one active energy component, could be used to determine the optimal total energy. However, generally, an SRAM is required to retain its data for an arbitrary length of time unrelated to its own access-delay. Consequently, the data-retention period cannot be parameterized by the access-delay, and a new parameter must be introduced to represent the total length of time data is retained. Specifically, idle data-retention consumes power, and to analyze its energy, the period of the retention-cycle, $T_{CYC,RTN}$, must be considered. Accordingly, $T_{CYC,RTN}$ corresponds to the average duration of time within which a required number of accesses are to be completed. The required number of accsses are designed as N. The data stored in the SRAM at the end of $T_{CYC,RTN}$ must correspond to these accesses, serving as the initial state for the subsequent set of accesses.

The actual length of time required to complete the N accesses can be set freely to optimize energy as long as it is less than $T_{CYC,RTN}$. This time to complete the accesses is designated as the access-period, T_{ACC}. For the remainder of the retention-cycle (i.e. $T_{CYC,RTN} - T_{ACC}$) only idle-data-storage is required. As discussed in detail in Section 2.2, the operating metrics associated with idle-data-retention are far less stringent than those associated with active data reads and writes. As a result, during idle-data-retention, the power can be much more aggressively reduced. The timing parameters relevant to SRAM energy are summarized in Figure 2-3.

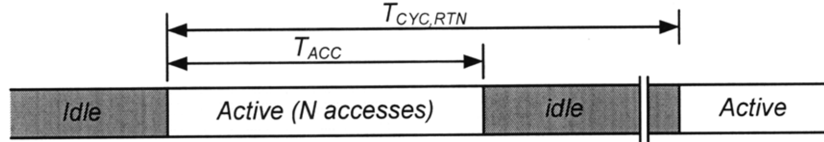


Figure 2-3: Summary of parameters relevant to SRAM energy.

A straight-forward and highly effective implementation of the low-energy data-retention mode involves dynamically reducing the voltage across the bit-cell array. This reduces the leakage-current by alleviating drain induced barrier lowering (DIBL), an increasingly prominent effect in advanced technologies. DIBL pertains to an effective decrease in the threshold-voltage brought on by increasing the MOSFET V_{GS}; large V_{GS} induces encroachment of the source/drain depletion regions into the channel region, reducing the gate to bulk biasing required for channel inversion.

Figure 2-4 shows the normalized leakage-current with respect to supply-voltage scaling, which also sets the V_{DS} of the devices. Predicite models have been used for this simulation, and as shown, well over an order of magnitude reduction in leakage-current can easily be achieved. The leakage-power savings further benefit from the supply-voltage reduction, leading to over 100x savings with 45nm CMOS when V_{DD} is scaled from 1.2V to 0.3V.

Practically, this approach has been successful by both reducing V_{DD} [52][53][54][45] and raising V_{SS} [55][56][57][58][59]. It should be mentioned that an additional approach involves reverse body-biasing to further reduce the leakage-current [60][61]. Nonetheless, the biasing employed in all of these cases can only be applied to the point where the data-storage margin is violated. Hence, the data-retention-voltage (V_{DRV}) is introduced in [45] to characterize the minimum V_{DD} at which data can reliably be retained by the bit-cells. As discussed further in Section 2.2, however, V_{DRV} is highly subject to variation. Consequently, closed-loop replica techniques have been employed to estimate the V_{DRV} limit dynamically, so that maximum idle-mode energy savings can be achieved [62]. In order to enforce a desired V_{DD} or V_{SS} voltage for the sub-array (i.e. V_{DDSUB} or V_{SSSUB}) during the idle-mode, the supporting circuits shown in Figure 2-5 have been used [63][64].

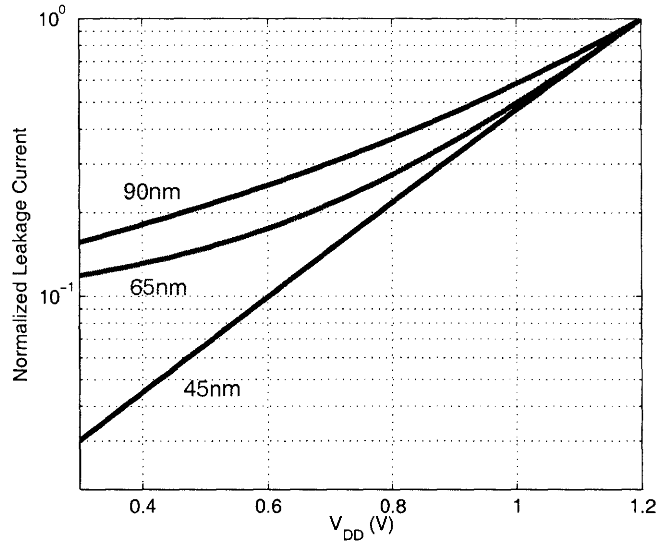


Figure 2-4: Normalized leakage-current reduction with respect to supply voltage for minimum-sized 90nm, 65nm, and 45nm devices due to DIBL (predictive models used).

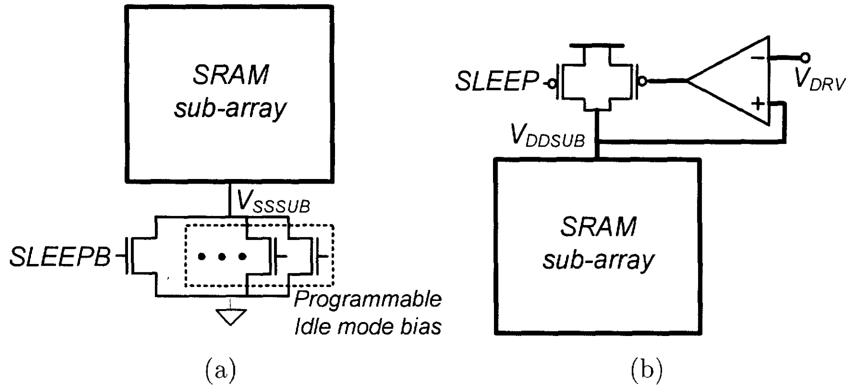


Figure 2-5: Circuitry to enforce idle-mode biasing using (a) programmable sleep switches [63] and (b) an operational-amplifier [64].

Regardless of the choice of the idle-mode biasing or the circuitry used to enforce it, it is critical that transitions between idle- and active-modes be made without compromising the biasing required in order to maintain the stringent active-mode operating margins. Consequently, careful signal timing is required to deriving the idle-mode $SLEEP$ signal which actuates the idle-mode biasing. Figure 2-6 shows an example of this signaling. In this case, full-cycle and half-cycle latencies corresponding to idle-to-active and active-to-idle transitions are inserted to ensure the corresponding operating margins are not violated [57].

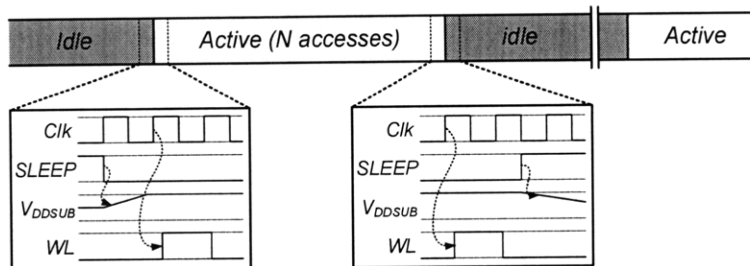


Figure 2-6: Waveforms corresponding to idle-to-active and active-to-idle mode transitions.

2.1.2 SRAM Sub-Array Optimal Energy

In this section, the average energy of an SRAM sub-array is considered, and more specifically, how it can be minimized by judicious selection of supply-voltage, V_{DD}, and device threshold-voltage, V_t, is analyzed. A typical SRAM is composed of many tiled sub-arrays, themselves consisting of a bit-cell array and access-control drivers/sensors. Additionally, global decoding and interfacing circuitry is also required. However, due to their very specific energy, performance, and operating characteristics (described above and further in Section 2.2), sub-arrays often employ a separate V_{DD} [65] and specialized devices [66], where the V_t is engineered for optimal operation. Because the sub-array critically determines the energy and performance of the entire SRAM, and because it offers independent control of V_{DD} and V_t, this section focuses on how the sub-array's energy can be optimized independently of the global decoding and interfacing circuitry.

Energy Components

Based on the operating model considered in Section 2.1.1, total sub-array energy, E_{TOT}, has four components, as indicated in Equation 2.1:

$$E_{TOT} = E_{ACC} + E_{LKG} + E_{IDL} + E_{OH} \tag{2.1}$$

The active-access-energy (E_{ACC}) and the leakage-access-energy (E_{LKG}) pertain to the active mode. E_{ACC} corresponds the switching energy required to perform reads and

writes, and E_{LKG} corresponds to the leakage-energy imposed by applying a supply-voltage across the array that must be large enough to ensure reliable reads and writes. The idle-data-retention energy (E_{IDL}) corresponds to data storage during the idle-mode, and it will also be referred to as the idle-mode energy. Finally, the overhead-energy (E_{OH}) corresponds to the overhead incurred due to altering the sub-array's biasing in accordance with idle-mode power reduction. These components are summarized in Figure 2-7, and they are described in more detail below.

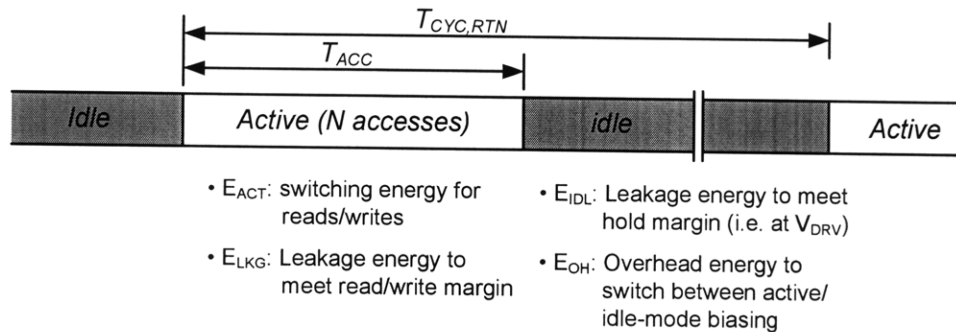


Figure 2-7: Summary of SRAM energy components.

(1) **Active-Access-Energy (E_{ACC}).** This represents the energy required to switch capacitive nodes in order to generate the control and data signals required to read and write bit-cells. Signal nodes that transition over the full-range from V_{DD} to ground require an active access-energy given by CV_{DD}^2, where C is the node capacitance. Full-swing signals typically include the one-hot enabled word-line, WL, for row selection, and the one-hot enabled column-select, $cSEL$, for multiplexed column selection in a column-interleaved array [67]. Of course, the internal nodes of the sense-amplifiers also switch from V_{DD} to ground. In total, the number of sense-amplifiers is equal to the number of columns in the sub-array divided by the column-multiplexing ratio, m.

The most significant source of active-access-energy consumption, however, is the bit-lines, BL, which are used to convey the stored read-data to the sense-amplifiers and to drive new write-data into the bit-cells. However, in some implementations, the BLs may not discharge completely during data-sensing.

Strictly speaking, to resolve the read-data, the BLs need only discharge to the required sense-amplifier input margin, V_{SNS}, which can be less than 100mV. Nonetheless, in practice, the BLs are often discharged beyond the sensing-margin to reduce the probability of data-disruption caused by sustained pulling of the bit-cell storages nodes towards the BL voltage near V_{DD}. During read-accesses, for instance, the design in [68] actively amplifies the signal on all BLs to full logic levels in order to avoid data-disruption. Accordingly, the total active-access-energy for reads of an $i \times j$ (i.e. i-column, j-row) sub-array is given by Equation 2.2, where the strong dependence on supply-voltage is clear:

$$E_{ACC,RD} = C_{WL}V_{DD}^2 + C_{cSEL}V_{DD}^2 + \frac{i}{m}C_{SA}V_{DD}^2 + iC_{BL}V_{DD}V_{SNS} \quad (2.2)$$

Similarly, the total active-access-energy for writes is approximately given by Equation 2.3:

$$E_{ACC,WR} = C_{WL}V_{DD}^2 + C_{cSEL}V_{DD}^2 + \frac{i}{m}C_{BL}V_{DD}^2 + i\frac{m-1}{m}C_{BL}V_{DD}V_{SNS} \quad (2.3)$$

(2) **Leakage-Access-Energy (E_{LKG}).** This represents the static energy consumed, even in the absence of active-accesses, just to generate a voltage across the sub-array that ensures the operating margins associated with active-accesses are reliably met. It comes about as a result of sub-threshold (and other) leakage-currents through the bit-cell devices that multiply with the supply-voltage, thereby consuming leakage-power. Since this source leads to static power dissipation, it must be integrated over a time interval to derive its energy. Minimally, the length of time that must be considered is T_{ACC}, the period required to complete some set number of accesses, N. Beyond this, the bit-cell biasing conditions no longer need to support the active-access operating margins, and biasing more conducive to minimum power-consumption can be enabled. Accordingly, the leakage-access-energy for an $i \times j$ sub-array is given by Equation

2.4, where it is assumed that the entire sub-array is biased with a single V_{DD} that must meet the active-access operating margins:

$$E_{LKG} = ij \int_{T_{ACC}} I_{LKG,BC} V_{DD} dt = ij I_{LKG,BC} V_{DD} T_{ACC} \tag{2.4}$$

In this expression, the dependence on V_{DD} is explicit through multiplication with the bit-cell leakage-current, which leads to the leakage-power. However, the dependence on V_{DD} is also implicit in two other ways: (1) the effect of V_{DD} on $I_{LKG,BC}$ through DIBL, and (2) the effect of V_{DD} on T_{ACC} through the V_{GS} available in order to generate bit-cell drive-current needed to discharge the BLs during data-sensing. Similarly, the dependence on threshold-voltage, V_t, is also implicit in two ways: (1) the effect of V_t on $I_{LKG,BC}$ through the sub-threshold current equation [69], and (2) the effect of V_t on T_{ACC} through the gate-overdrive (i.e. $V_{GS} - V_t$) necessary to generate bit-cell drive-current. Additionally, V_t also affects the ability of the bit-cells to meet the operating margins given a particular V_{DD}. Consequently, as described in Section 2.2, V_t has a direct effect on the minimum V_{DD} allowed.

(3) **Idle-Data-Retention Energy (E_{IDL}).** This represents the static energy required to retain the data, without any active-accesses, until the end of some required period. Considering the power-management scenario described in Section 2.1.1, system operations will require an average number of accesses, N, every $T_{CYC,RTN}$ seconds. The operating point of the sub-array may be chosen to optimize energy as long as the N accesses are completed in a period less than $T_{CYC,RTN}$. For the remainder of the time until the end of $T_{CYC,RTN}$, however, the data must be retained so that it is available for the next set of accesses. This cycle is shown in Figure 2-7. Accordingly, the idle-data-retention energy is given by Equation 2.5:

$$E_{IDL} = ij \int_{T_{CYC,RTN}-T_{ACC}} I_{DRV,BC} V_{DRV} dt = ij I_{DRV,BC} V_{DRV} (T_{CYC,RTN} - T_{ACC}) \quad (2.5)$$

Here, V_{DRV} refers to the data-retention voltage [45], and $I_{DRV,BC}$ refers to the leakage-current of the bit-cell at V_{DRV}. In this expression, the dependence on V_t is implicit since it affects $I_{DRV,BC}$ through the sub-threshold current equation. Further, as described in Section 2.2, V_t also affects the minimum V_{DRV} achievable. Although it is possible to adjust V_t dynamically [60][61] in order to optimize the idle-mode energy, compared to V_{DD} such adjustments are more difficult to make over an aggressive range. Finally, as mentioned previously, both V_{DD} and V_t affect T_{ACC}.

(4) **Overhead Energy (E_{OH}).** This represents the energy consumed in order to transition to the low-energy idle-mode state. During the idle-mode, the array must be rebiased by changing V_{DD}, V_{SS}, and/or the body-bias. This involves appropriately charging the supply, ground, or back-gate capacitance for the entire array. For the case of changing the sub-array supply-voltage from V_{DD} to V_{DRV}, the overhead energy, which is consumed once every $T_{CYC,RTN}$, is given by Equation 2.6, where C_{VDD} is the total power-supply capacitance:

$$E_{OH} = C_{VDD} V_{DD} (V_{DD} - V_{DRV}) \quad (2.6)$$

In this expression, the dependence on V_{DD} is explicit, and the dependence on V_t, which limits the minimum achievable V_{DRV} as mentioned above, is implicit. It should be noted that some finite time is required in order to ensure complete transition between the idle-mode and active-mode biasing, and it is critical to consider this in order to avoid violating the different operating margins associated with each mode. Nonetheless, the leakage-energy that is consumed during the transition period is relatively insignificant, since C_{VDD} is typically very large (i.e. >100pF) and the transition time required is on the order of only a few

clock-cycles [57]. Finally, since E_{OH} is an unavoidable overhead associated with transitioning to the low-energy idle-mode, it is useful to analyze whether the energy savings yielded will be sufficient to exceed the energy overhead. Minimally, this requires that [70]

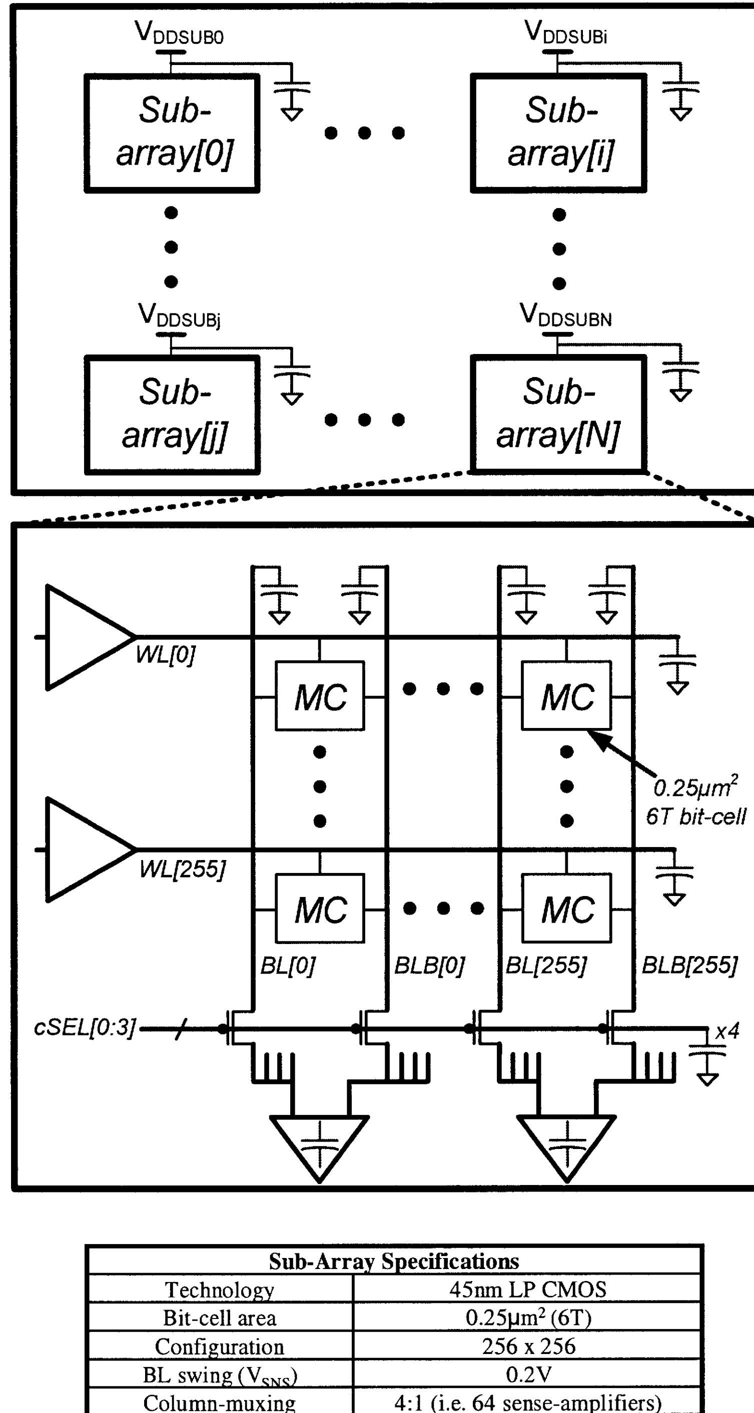
$$E_{OH} < ij(I_{LKG,BC}V_{DD} - I_{DRV,BC}V_{DRV})(T_{CYC,RTN} - T_{ACC}), \qquad (2.7)$$

and even further, the overhead associated with circuitry to support the rebiasing must also be considered.

Sub-Array Energy Analysis

To ascertain V_{DD} and V_t targets that lead to optimal sub-array energy, a practical case for a low-power high-density SRAM is considered. The specifications of the sub-array are shown in Figure 2-8. In particular, an LP 45nm CMOS technology is used. The sub-array consists of 256 columns and 256 rows of bit-cells that have been designed to occupy a layout area of $0.25\mu m^2$ using actual SRAM design-rules for the technology. Column-multiplexing of 4:1 is assumed, such that 64 (out of the 256) cells are accessed each cycle. Layout extraction is performed to determine the parasitic capacitances of the word-lines (WL), bit-lines (BL), column-select-lines ($cSEL$), sense-amplifiers, and power-supply (V_{DDSUB}, which will be referred to as V_{DD} for the remainder of this analysis). Finally, the total voltage-swing on the bit-lines is assumed to be 200mV during read-accesses. All other digital control signals are assumed to be full-swing, from ground to V_{DD}.

To characterize the energy, simulations are performed by scaling V_{DD} for the entire sub-array and scaling V_t of the bit-cell devices. This is achieved by adjusting the $VTH0$ parameter of the $BSIM4$ transistor models, which corresponds to the threshold voltage of a long-channel device with zero substrate bias [71]. The effects of device variations, and how they scale with V_{DD} and V_t are not considered here. Instead, the optimal targets are being established. The impact variation has on parameters relevant to the energy will be considered in Section 2.3, by revising the energy analysis.



Sub-Array Specifications	
Technology	45nm LP CMOS
Bit-cell area	$0.25\mu m^2$ (6T)
Configuration	256 x 256
BL swing (V_{SNS})	0.2V
Column-muxing	4:1 (i.e. 64 sense-amplifiers)

Figure 2-8: Sub-array specifications for energy analysis.

In particular, the data-retention voltage, V_{DRV}, which, in the presence of variation, is heavily dependant on V_t, will be taken to equal 0.4V for the initial analysis.

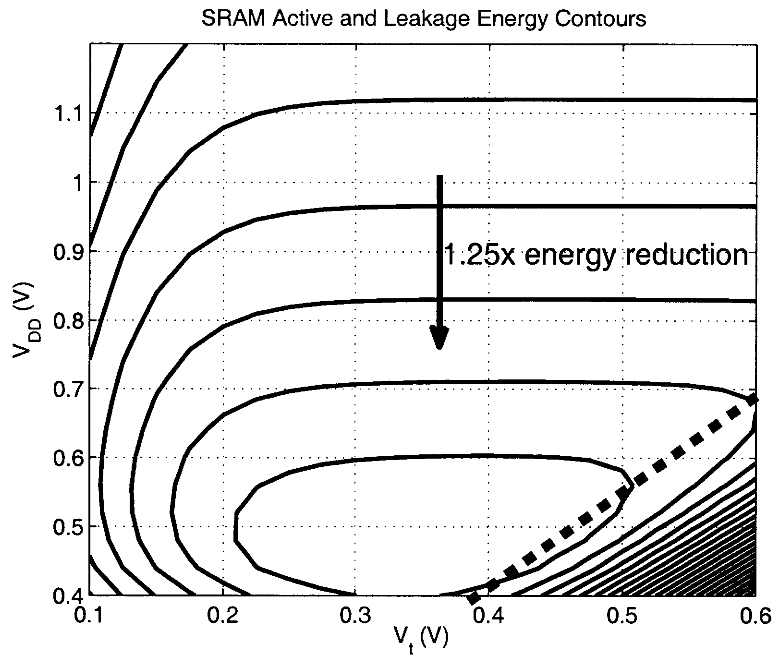
The average number of accesses (N) required for logical operations, and the aver-

age time required to complete a logical operation ($T_{CYC,RTN}$) are application dependant parameters that can significantly affect the optimal total energy of the sub-array. For instance, as $T_{CYC,RTN}$ becomes very long, the leakage-energy, specifically during the idle-mode (i.e. E_{IDL}), dominates over all of the other components, and it largely negates the impact of V_{DD} all together. However, for most of the low-power applications discussed in Chapter 1, the time-scales of interest lead to a dependence on all of the energy components. To proceed with the analysis, N is assumed to be 1024, which corresponds to an access of every bit-cell in the 64kb sub-array (since 64 cells are accessed each cycle). Additionally, $T_{CYC,RTN}$ is set to $10ms$, $1ms$, $100\mu s$, and $10\mu s$ to consider various performance constraints. For the array configuration considered, E_{IDL} overwhelmingly dominates when $T_{CYC,RTN}$ is much longer than $10ms$.

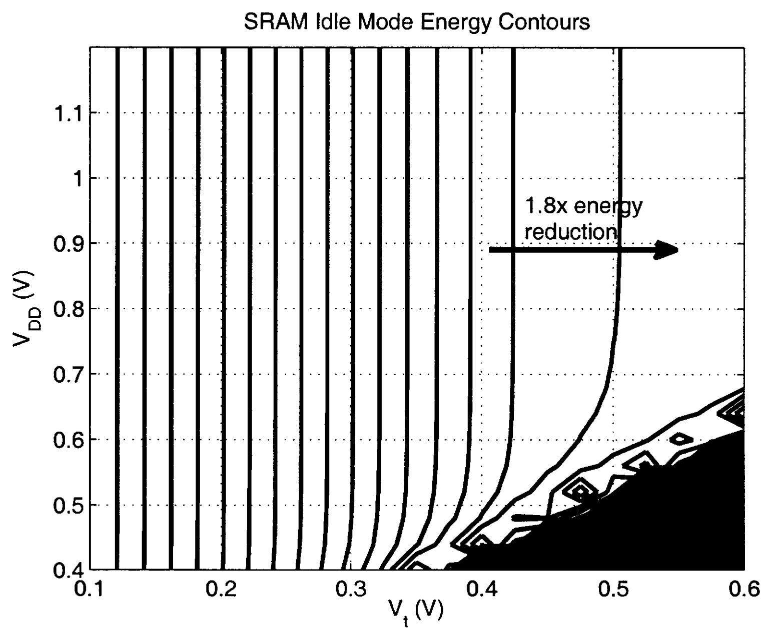
Before analyzing the total energy, the energy components are discussed. Figure 2-9 shows the active-mode energy (corresponding to $E_{ACC} + E_{LKG}$), idle-mode energy, and overhead energy plotted as log-magnitude contours with respect to V_{DD} and V_t. Here, $T_{CYC,RTN}$ is set to $1ms$. As $T_{CYC,RTN}$ and N are varied, the trends observed for each component remain constant, but the relative magnitudes of the components change. For instance, large $T_{CYC,RTN}$ and small N elevates the importance of idle-mode energy with respect to active-mode energy; similarly, overhead energy has reduced prominence as N increases, since it gets amortized over more active-accesses.

The contours observed for active-accesses (Figure 2-9a) are typical for digital circuits [44]. At low V_{DD} (0.4-0.6V), the sub-array speed is significantly reduced, so minimizing the leakage currents, by increasing V_t from 0.1-0.3V, favorably affects the energy; at higher V_{DD}, the energy is overwhelmingly dominated by capacitive switching. As V_t is increased beyond 0.4V (in the region below the dotted line of Figure 2-9a), deep sub-threshold operation leads to compromised logic levels causing artifacts leading to increased energy. For the considered array configuration and technology, the active-mode energy points to an optimal V_{DD} and V_t of approximately 0.5V and 0.35V, respectively.

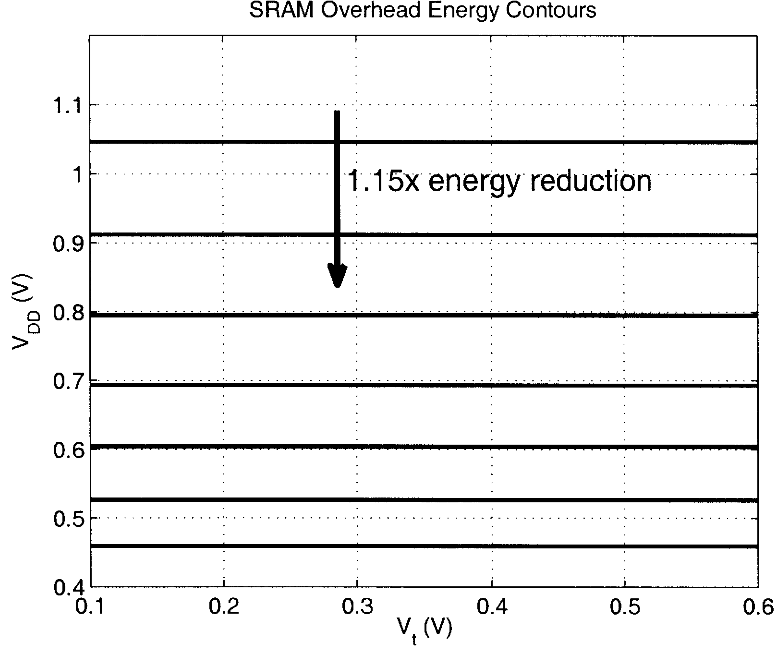
As expected, the idle-mode energy (Figure 2-9b) is strongly dependant on V_t, due



(a) Active-mode (switching and leakage) energy components



(b) Idle-mode leakage-energy component



(c) Idle-active transition overhead energy

Figure 2-9: Sub-array individual energy components.

to the effect of V_t on sub-threshold leakage-currents. Here, the blocked-out portion corresponds to a region where the performance constraint (specified by $T_{CYC,RTN}$ and N) is not met due to excessive access-delay, T_{ACC} (i.e. $T_{ACC} > T_{CYC,RTN}$). At the boundary of this region, the idle-mode energy degenerates to zero since the entire data-retention period ($T_{CYC,RTN}$) is spent in the active-mode, and the irregular contours represent artifacts due to the limited resolution of the plotted points.

In Figure 2-9b, the benefit of increasing V_t beyond 0.4V degrades for high V_{DD} (as indicated by the increasing distance between the energy contours). Here, leakage-power sources other than sub-threshold currents, namely gate- and junction-currents, start to become significant. At lower V_{DD} (i.e. below 0.6V) and high V_t (i.e. above 0.3V), the idle-mode energy seems to rapidly decrease as the threshold-voltage in increased (as indicated by the contours tapering together). In this region, which corresponds to sub-threshold operation, the active-mode access-delay (T_{ACC}) increases rapidly with V_t. As a result, a smaller total portion of $T_{CYC,RTN}$ is spent in the idle-mode, and the idle-mode energy appears to decrease quickly.

The overhead energy (Figure 2-9c) represents the cost of charging the sub-array

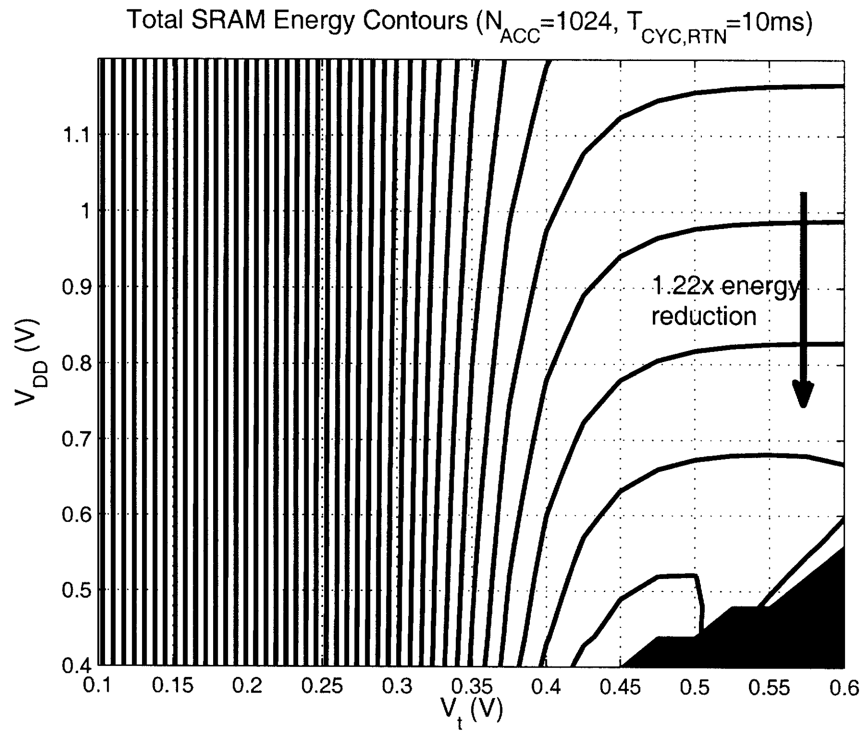
power-supply node between the active-mode voltage and the data-retention voltage. Consequently, it has a straight-forward dependence on V_{DD}.

Combining all of the energy components, the total sub-array energy is plotted as log-magnitude contours with respect to V_{DD} and V_t in Figure 2-10. Once again, the blocked region represent supply- and threshold-voltages where the performance constraint is not met (i.e. $T_{ACC} > T_{CYC,RTN}$).

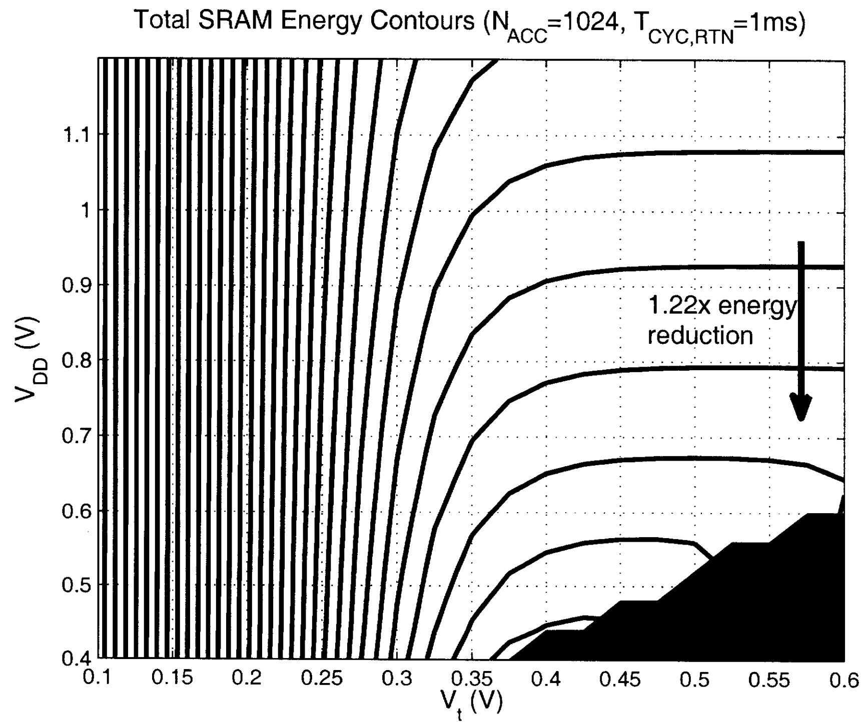
Several important observations can be drawn from these plots. First, compared to the active-mode energy (shown in Figure 2-9a), which follows the behavior of general digital logic [44], the total energy favors a lower supply-voltage. For instance, in all cases a V_{DD} less than 0.5V is optimal, and, in fact, even lower supply-voltages would be preferable if the performance constraint could be met (this result is discussed further below). The preference towards low supply-voltages occurs since, in the absence of long-term data-retention, the leakage-currents can be completely negated at the end of T_{ACC}, which is shortened greatly by raising V_{DD}. However, the need for data-retention precludes complete leakage-current gating, somewhat attenuating the benefit of raising V_{DD}.

Second, as mentioned, in all cases, the energy contours continue to decrease fairly rapidly into the regime where the performance constraint is not met. The important consequence of this is that any means to improve the performance will enable further V_{DD} and V_t scaling, leading to further reduction in the sub-array energy.

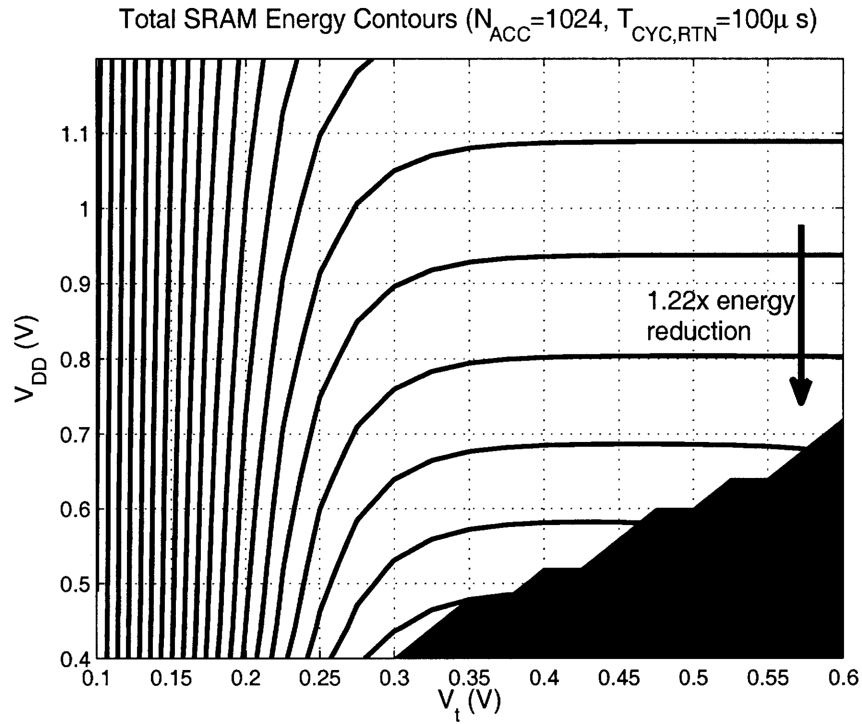
An additional result, which must be qualified, seems to be that performance limitation in this manner indicates no need for the idle-mode at all; in particular, the sub-array should be operated with a relatively high threshold-voltage and the lowest possible supply-voltage required to just meet the $T_{CYC,RTN}$ constraint. Of course, in large SRAMs where several sub-arrays are tiled, the benefit of the idle-mode is clear, as it minimizes the leakage-power of all the inactive sub-arrays. However, even with very few sub-arrays, the idle-mode can be important. Section 2.3 will show that reducing the access-period beyond $T_{CYC,RTN}$ can be beneficial in order to reduced excessive active-mode leakage-energy; this requires raising the supply-voltage in order to overcome performance degradation from variation, allowing sooner tran-



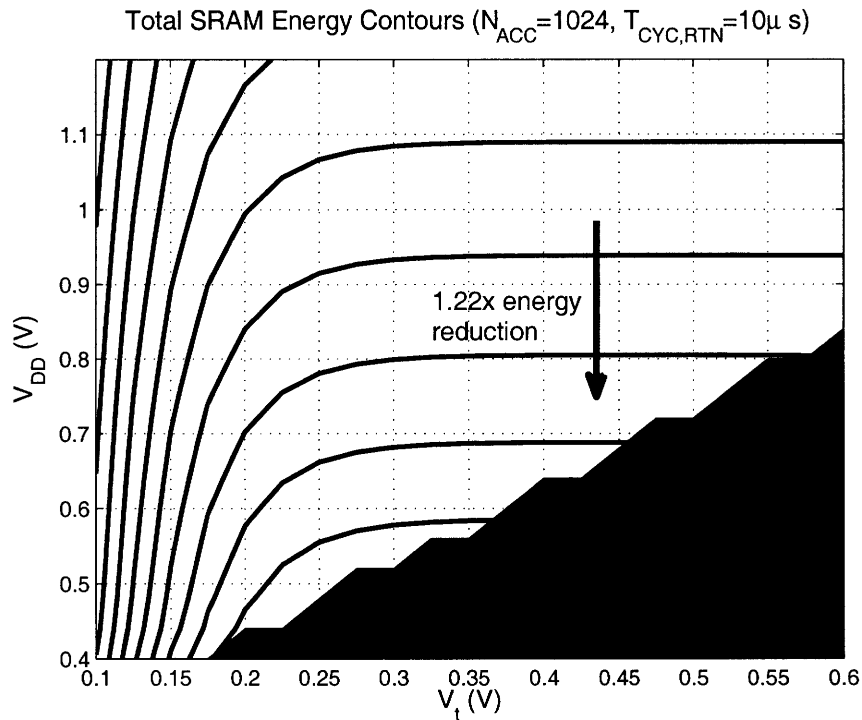
(a) Total energy for $T_{CYC,RTN} = 10ms$.



(b) Total energy for $T_{CYC,RTN} = 1ms$.



(c) Total energy for $T_{CYC,RTN} = 100\mu s$.



(d) Total energy for $T_{CYC,RTN} = 10\mu s$.

Figure 2-10: Sub-array total energy (at room temperature) for various performance requirements (specified by $T_{CYC,RTN}$).

sition into the low-power idle-mode. The idle-mode is also beneficial for recovering additional energy imposed by margining. For instance, some V_{DD} margin is necessary to support changes in the operating conditions and instantaneous peaks in the performance demands (which require shortening the access-period below the average $T_{CYC,RTN}$ considered). Figure 2-11 considers the effect of this margin if provisions for the idle-mode are not included. For the $T_{CYC,RTN} = 10ms$ case, the energy components (normalized to E_{TOT}) are shown along a slice corresponding to the $V_t = 0.45V$ axis (the optimal achievable energy occurs along this slice). An additional energy component, $E_{LKG,CYC,RTN}$, is also plotted, which corresponds to the leakage-energy that would be incurred if the active-mode V_{DD} were used for the entire duration of $T_{CYC,RTN}$. Although $E_{LKG,CYC,RTN}$ degenerates to E_{LKG} at the optimal V_{DD} (which is the minimum voltage of 0.42V), less than 0.15V of V_{DD} margin makes it the dominant source of energy, and it increases rapidly from there. Hence, the idle-mode provides a means to minimize the excess energy imposed by the required margining.

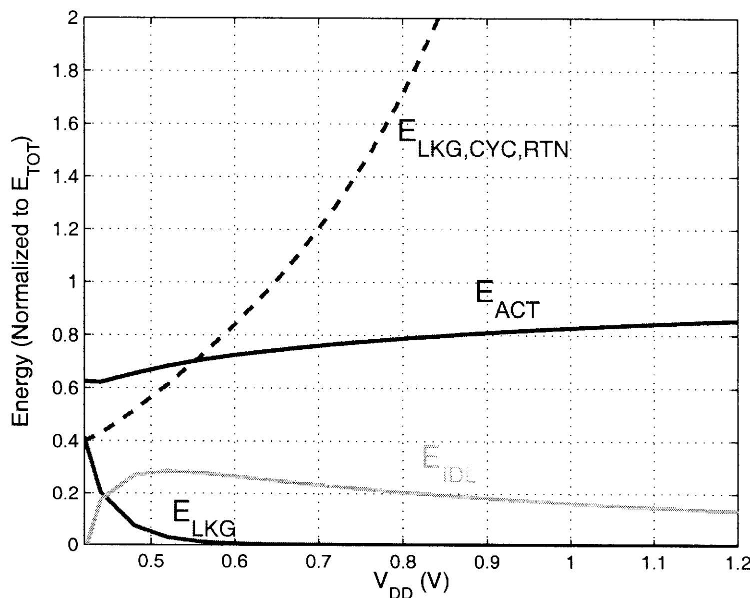


Figure 2-11: Energy components for $T_{CYC,RTN} = 10ms$ along $V_t = 0.45V$ axis.

2.2 SRAM Operating Margins and Metrics

Aside from the performance constraint specified by $T_{CYC,RTN}$, SRAMs must meet several other operating margins that are not considered in the analysis of Section 2.1.2. The optimal V_{DD} and V_t trends established there are only targets; but enabling actual sub-array operation at those targets requires overcoming the associated operational challenges. This section examines how read-margin, write-margin, data-retention, and read-current depend on V_{DD} and V_t, particularly in the presence of variation.

Generally speaking, the motivation to reduce V_{DD} and raise V_t, based on sub-array energy optimization, is opposed not only by the ensuing degradation of noise-margins, but also by an enhanced sensitivity of MOSFETs to variability. Due to the tendency towards large sub-arrays, the level of variation observed in SRAMs is extreme, typically beyond the 5σ level. Substantial effort is devoted to minimizing SRAM variation. For instance, at the device-level, implant doping (material and orientation) as well as layout features are carefully controlled [72]. Similarly, at the array-level, bit-cell redundancy is widely used to mitigate the impact of extreme variation [73]; nonetheless, in 256×256 sub-arrays, variation beyond the 4σ level can still be expected to limit operation [74].

Figure 2-12 shows the effect of variation on MOSFET drain-current (with $V_{GS} = V_{DS} = V_{DD}$) in two lights. In Figure 2-12a, an NMOS with $V_t = 0.3V$ is considered as V_{DD} is scaled for both a mean and 4σ device. The distance between the two widens drastically as supply-voltage is reduced (indicating a degrading ratio of mean-to-4σ current). This comes about due to the increasing dependence of the gate-overdrive, $V_{DD} - V_t$, on V_t fluctuations combined with an increasing dependence of the drain-current on that gate-overdrive (which ranges from linear to exponential towards the sub-threshold regime [75]).

In Figure 2-12b, an NMOS with V_{GS}=V_{DS}=1V is considered as V_t is scaled. Threshold-voltages that are engineered to be higher exhibit increased σV_t. This is due to the need to increase dopant concentration, which leads to more severe random dopant fluctuation (RDF) [76]. Consequently, to account for the effect that V_t scaling

has on variation, σV_t has been adjusted using the relationship of Equation 2.8 [77]:

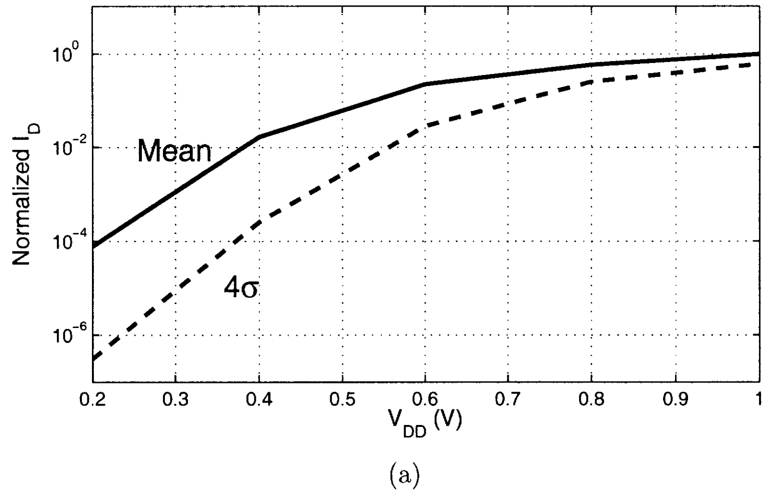
$$\begin{aligned}
\sigma V_t &\propto \sqrt{q^2 N_{SUB} W_{DEP}} \\
\sigma V_t &\propto \sqrt{\frac{q(V_t - V_{FB} - 2\phi_F)}{C_{OX}}} \\
\sigma V_t &\propto \sqrt{(V_t - V_{FB} - 2\phi_F)} \\
\sigma V_t &\propto \sqrt{V_t + 0.1}.
\end{aligned} \tag{2.8}$$

(Here, $-V_{FB} - 2\phi_F \approx 0.1$, which has been validated through several data-points from 65nm fabs [77].) As a result, with increasing V_t, the 4σ current deviates increasingly from the mean current, particularly as V_t approaches V_{DD}, tending towards an exponential impact in sub-threshold.

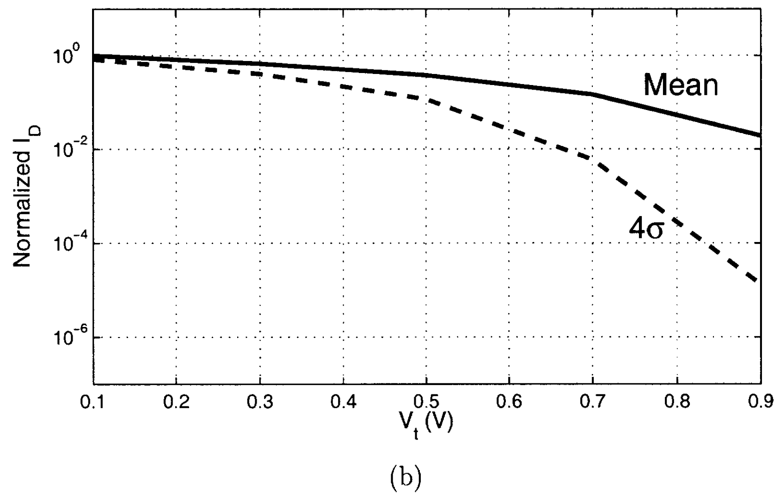
In the following subsections, Monte Carlo simulations are performed on the nominal process conditions, and the statistical device parameters affected by variation (including V_t), are sampled from a Gaussian distribution while considering the impact of V_{DD} and V_t scaling. Here, the effect of local-variation (i.e. intra-die) [41], which most prominently limits SRAM functionality [78], is combined with global (inter-die) process-skews in order to illustrate the total effect.

2.2.1 Read-Margin

The read SNM quantifies the extent to which a 6T bit-cell can reliably hold each of the two data states required while being subjected to a static read condition. The read SNM is illustrated graphically in the butterfly plots of Figure 2-13. Here, the transfer-functions between the bit-cell's data storage nodes (i.e. from NT-NC and from NC-NT) are superimposed. As shown in Figure 2-13a, the static read condition implies that the access-devices ($M5-6$) are enabled and the bit-lines are held at V_{DD}. The cell's ability to reliably hold both data states depends on the transfer-functions (plotted in Figure 2-13b) intersecting at two valid logic levels, and it is quantified by the length of the diagonal of the largest square embedded in the transfer-function



(a)



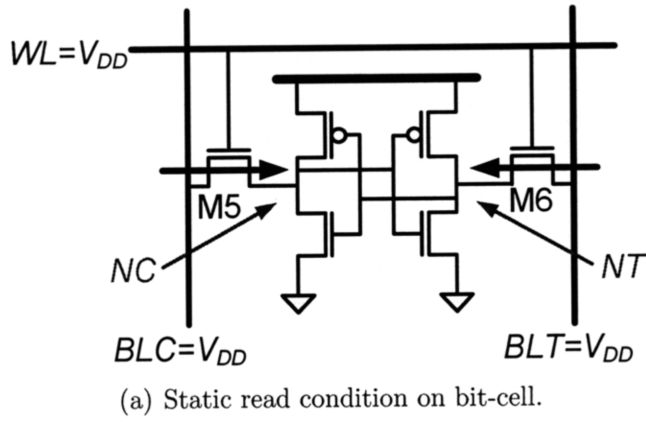
(b)

Figure 2-12: Mean and 4σ drain-current for minimum sized NMOS in 45nm CMOS with respect to (a) V_{DD} (with V_t=0.3V) and (b) V_t (with V_{DD}=1V).

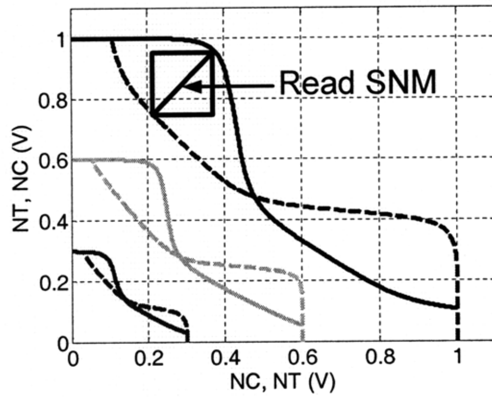
lobes [40].

Figure 2-13b shows how variation can shift the transfer-functions [79], and how supply-voltage scaling degrades the noise margin, easily leading to the loss of the read SNM. Similarly, threshold-voltage scaling, has a detrimental effect through the increase in σV_t it introduces.

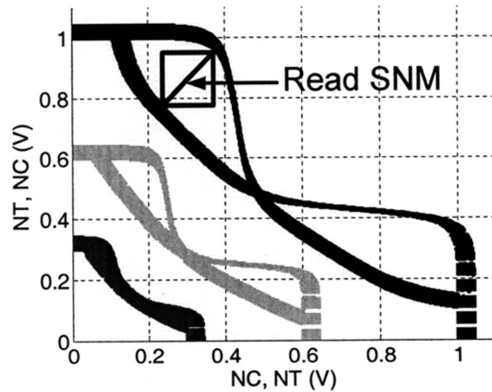
To determine the combined effect of supply- and threshold-voltage scaling, Figure 2-14 shows the mean and variation-affected read SNM with respect to V_{DD} and V_t. For the variation-affected case, 4σ local variation is considered on top of the process global-variation. As shown, variation strongly limits the region where read SNM is



(a) Static read condition on bit-cell.



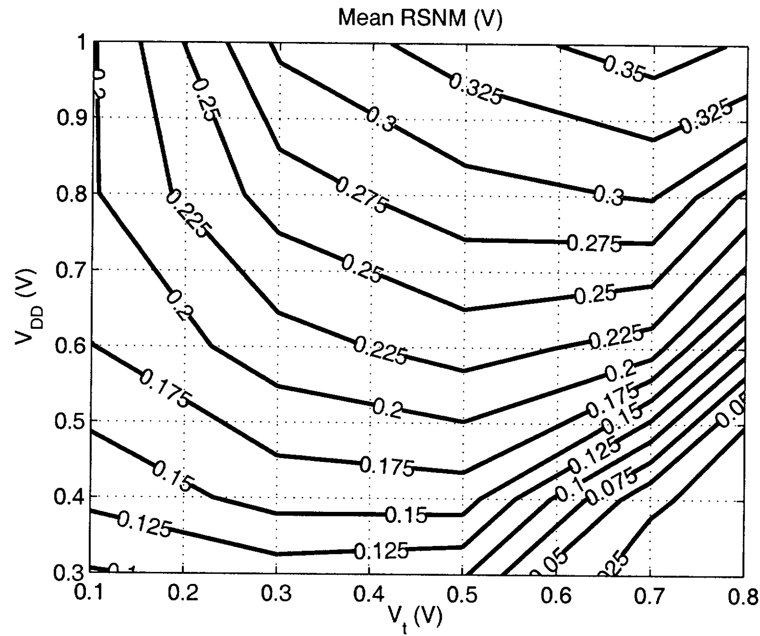
(b) Read butterfly plot with no variation.



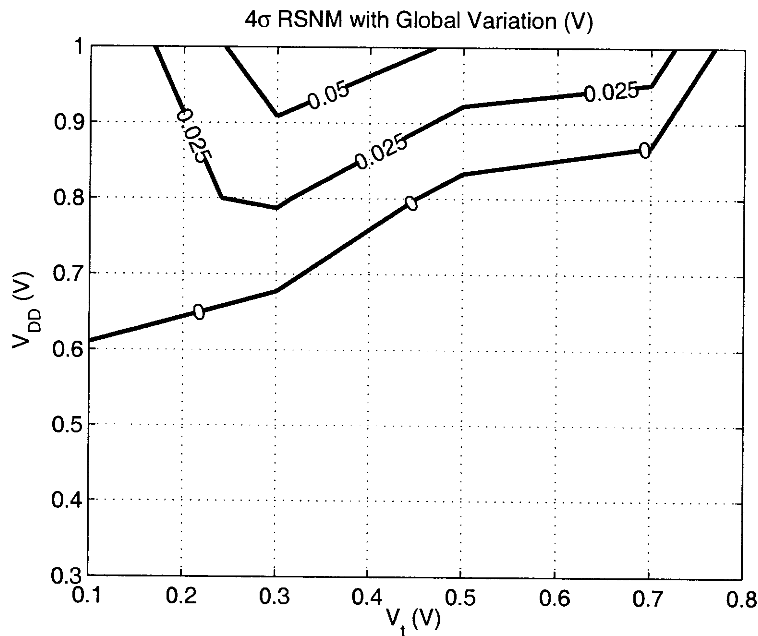
(c) Read butterfly plot with variation.

Figure 2-13: Read SNM definition through butterfly plots.

preserved, specifically restricting operation at low V_{DD} and high V_t, where sub-array energy tends to be optimized.



(a)



(b)

Figure 2-14: 45nm $0.25 \mu m^2$ bit-cell read SNM contours for (a) mean case, and (b) 4σ (on top of global variation) case.

2.2.2 Write-Margin

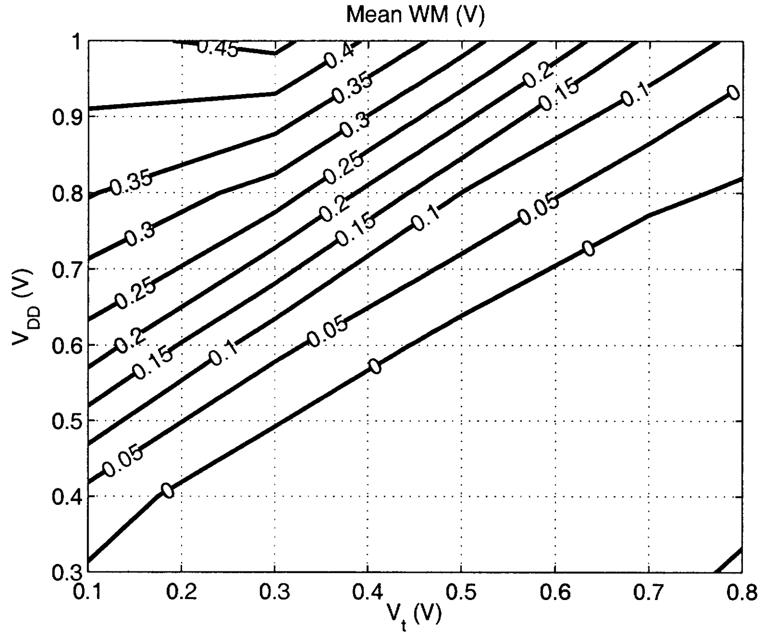
Several metrics exist to quantify write-ability [80][42]. One that relates well to the read-margin is the ability to make the bit-cell mono-stable at the logic state intended to be stored [81]. This corresponds to the negative of the read SNM, and is used here.

Figure 2-15 shows the mean and 4σ (on top of global variation) write-margin with respect to V_{DD} and V_t. Once again, variation strongly limits the functional region and specifically opposes that where sub-array energy is optimized. It should be noted, that the result shown is for a dense $0.25\mu m^2$ bit-cell, which severely constrains the sizing of the constituent devices. In a practical cell, however, threshold voltage engineering provides an additional means, beyond just sizing, to set the required relative device strengths. Here, in order to develop the general trends with minimal complexity, all V_t's are assumed to scale equivalently; however, in a practical cell, selective V_t engineering actually leads to better write-margin. Nonetheless, the increased impact of variation on write-margin seen at low V_{DD} and high V_t remains, and it critically contributes to limiting sub-array energy.

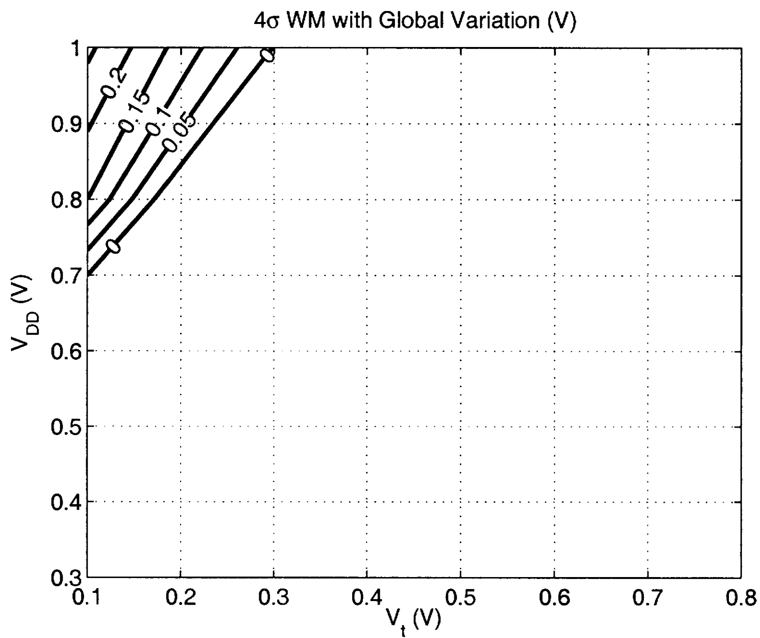
2.2.3 Hold-Margin (and Data-Retention-Voltage)

The hold-margin quantifies the ability of the bit-cell to idly retain data in the absence of read or write conditions. The hold SNM [40] is analogous to the read SNM; however, as shown in Figure 2-16a, it implies that the bit-cell access-devices are disabled, precluding the disruption of the storage nodes NT/NC by the bit-lines near V_{DD}. Consequently, as shown in Figure 2-16b, the hold SNM can be significantly larger than the read SNM. As shown in Figure 2-16c, this implies the possibility of low V_{DD} (or high V_t) data-retention even in the presence of variation, leading to much lower power consumption.

In this manner, the hold-margin is directly related to the data-retention voltage. Here, the hold SNM is used as the hold-margin, and the V_{DD} where it equals zero (in the 4σ on top of global-variation case) is taken to be the data-retention voltage, V_{DRV}; of course, in practice it is prudent to also introduce some additional engineering



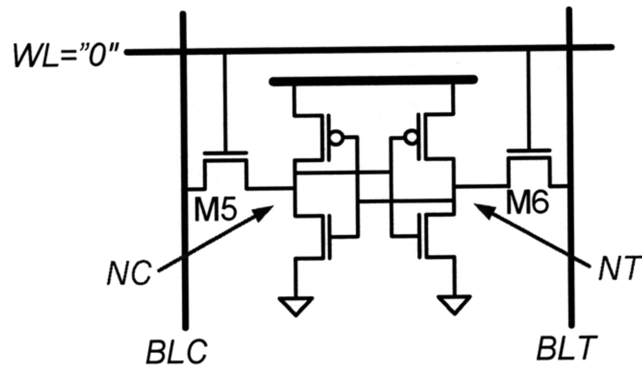
(a)



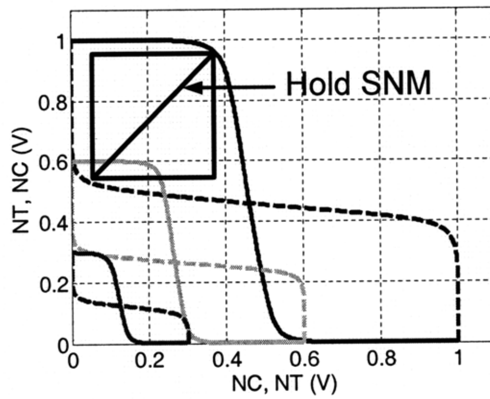
(b)

Figure 2-15: 45nm $0.25\mu m^2$ bit-cell write-margin contours for (a) mean case, and (b) 4σ (on top of global variation) case.

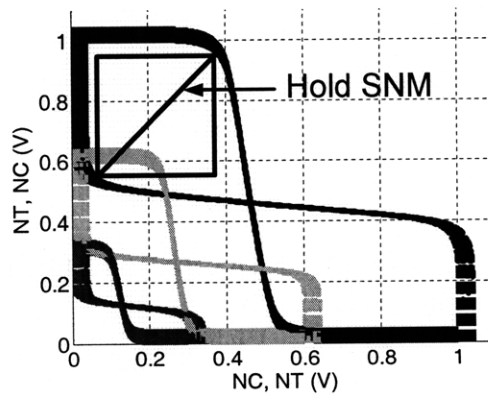
margin when setting the idle-mode V_{DD}, but the additional margin, which degrades the power-savings, can be minimized if V_{DRV} can be accurately determined either through simulation [82][83] or run-time sensing and estimation [62].



(a) Static hold condition on bit-cell.



(b) Hold butterfly plot with no variation.



(c) Hold butterfly plot with variation.

Figure 2-16: Hold SNM definition through butterfly plots.

Figure 2-17 shows the mean and 4σ (on top of global-variation) hold SNM with respect to V_{DD} and V_t. Due to higher dependence on V_t fluctuations and elevated σV_t, increasing V_t tends to raise the minimum tolerable V_{DRV}. Consequently, the favorable impact of reducing the leakage-current degrades as V_t is increased. This effect is particularly important at very high threshold-voltages, where the leakage-current is dominated by gate and junction sources; here, V_t scaling can actually increase the idle-mode energy due to the higher V_{DRV} required.

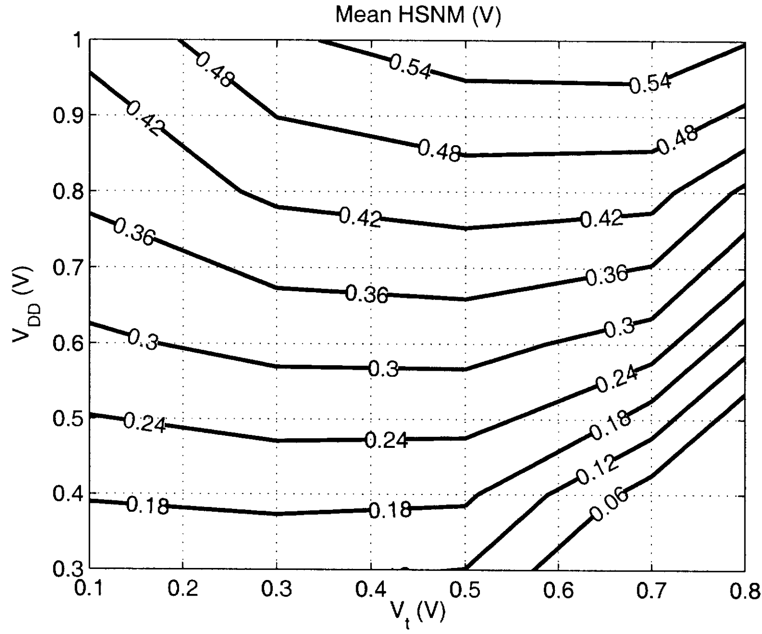
2.2.4 Cell Read-Current

The read-current, I_{RD}, is the current sunk by the bit-cell from the bit-line immediately after its access devices are enabled. The biasing condition implied here is that the bit-lines are at their precharge voltage, which is typically V_{DD}. The read-current is a critical metric for sub-array performance, and, as discussed in Section 2.1.2, it also strongly affects the minimum achievable energy.

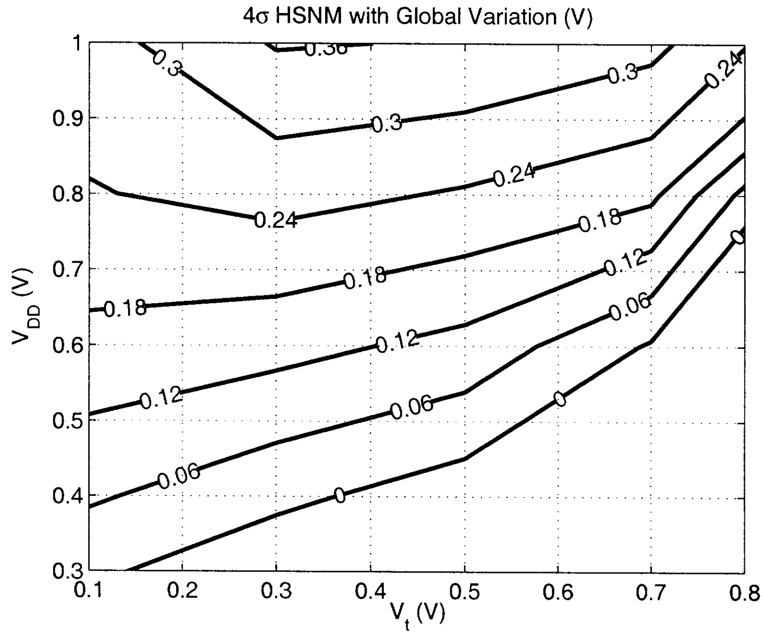
Figure 2-18 shows the mean and 4σ $log_{10}(I_{RD})$ with respect to V_{DD} and V_t. As expected, lowering V_{DD} and raising V_t strongly reduces the mean I_{RD} and increases the further degradation from variation. Improving cell-drive capability is critical for low-energy sub-arrays not only because this enables more aggressive V_{DD} and V_t scaling under set performance constraints, but also because it overcomes functionality failures that are fundamental to SRAMs at the low-energy operating points. These failures are further discussed in Chapter 3.

2.3 SRAM Energy with Variation

Since V_{DD} and V_t scaling so severely elevates the effect of device variation, the optimal energy analysis of section Section 2.1.2 must be revised. In particular, three important effects emerge: (1) the access-period, T_{ACC} is much longer, (2) the minimum achievable V_{DRV} is higher, and (3) the total aggregate leakage-current is higher due to the variation gain factor (illustrated in Figure 2-1). The resulting impact on the total energy is considered below.



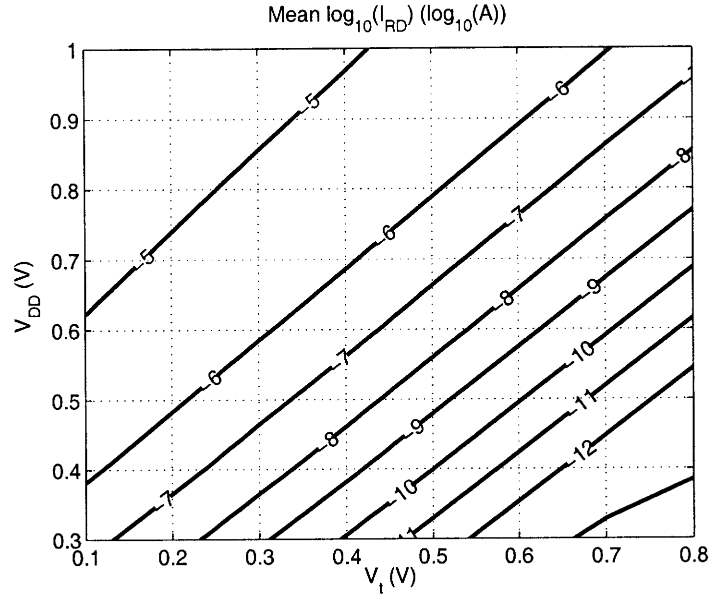
(a)



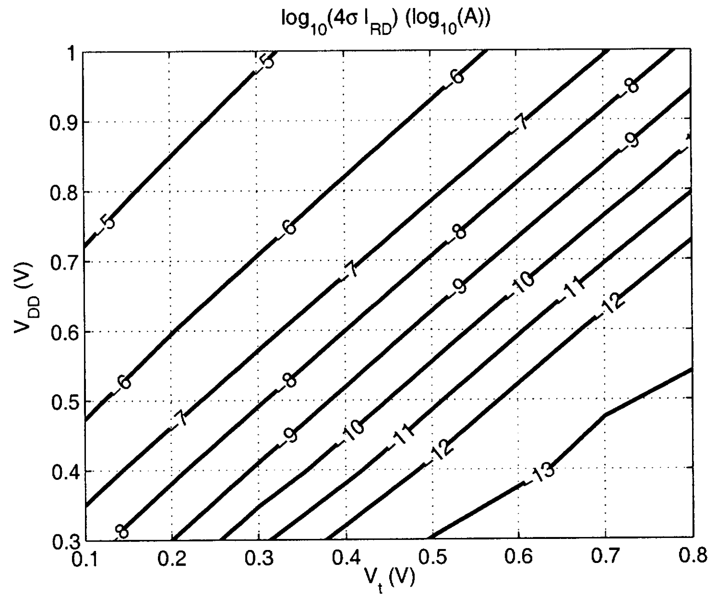
(b)

Figure 2-17: 45nm $0.25\mu m^2$ bit-cell hold SNM contours for (a) mean case, and (b) 4σ (on top of global variation) case.

First, regarding the increase in T_{ACC}, the most important implication is that the supply- and threshold-voltage region where the performance constraint is not met is significantly expanded. Hence, the energy optimization achievable through V_{DD}



(a)

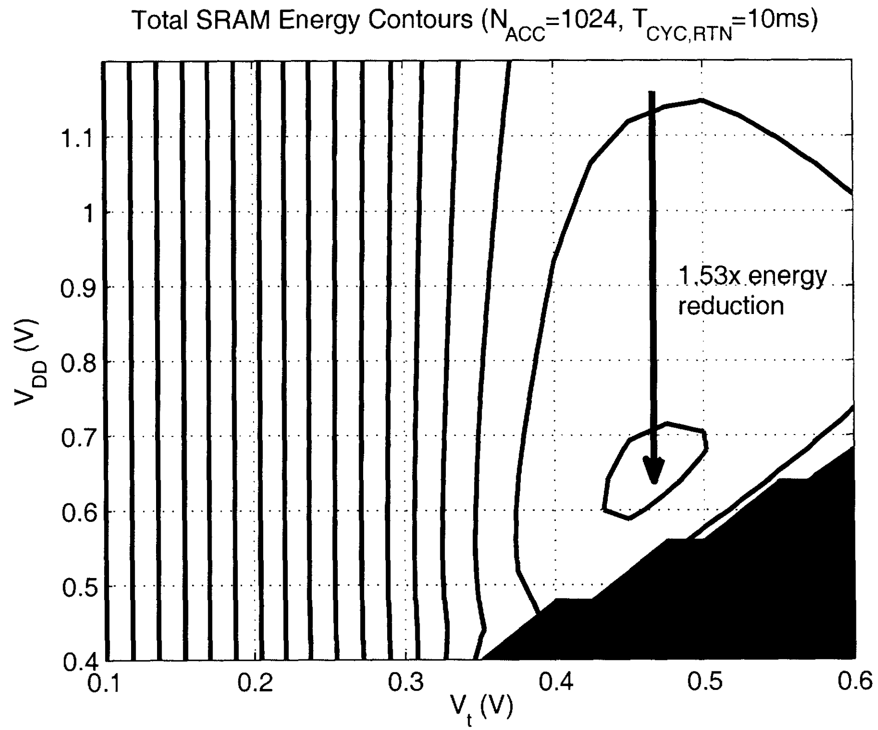


(b)

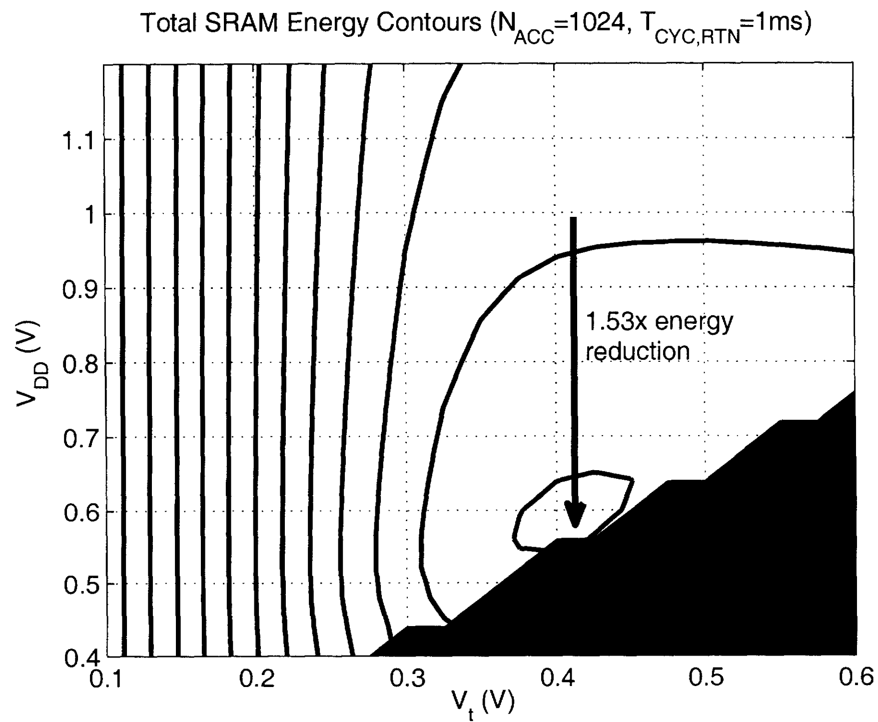
Figure 2-18: 45nm $0.25\mu m^2$ bit-cell read-current contours (log-magnitude) for (a) mean case (b) 4σ.

and V_t scaling is directly limited. As shown in Figure 2-19c-d, this, is particularly detrimental in the high performance cases (i.e. $T_{CYC,RTN} = 100\mu s, 10\mu s$).

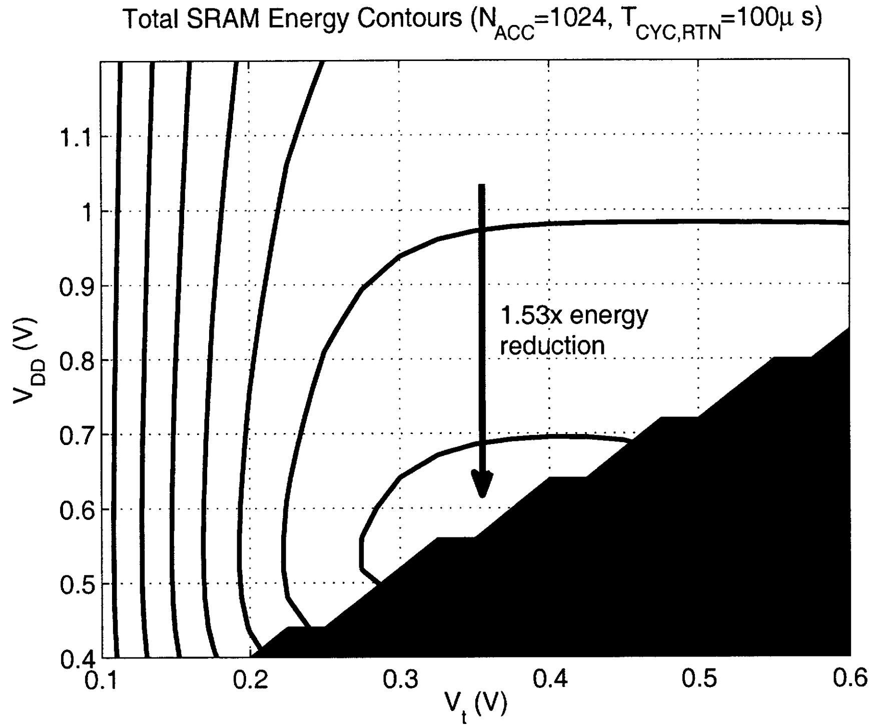
Second, for the low performance cases, the optimal energy is not limited by the performance constraint, $T_{CYC,RTN}$. For instance for the $T_{CYC,RTN} = 10ms$ case (shown in Figure 2-19a), the optimal point is $V_{DD} = 0.65V$ and $V_t = 0.475V$, which oc-



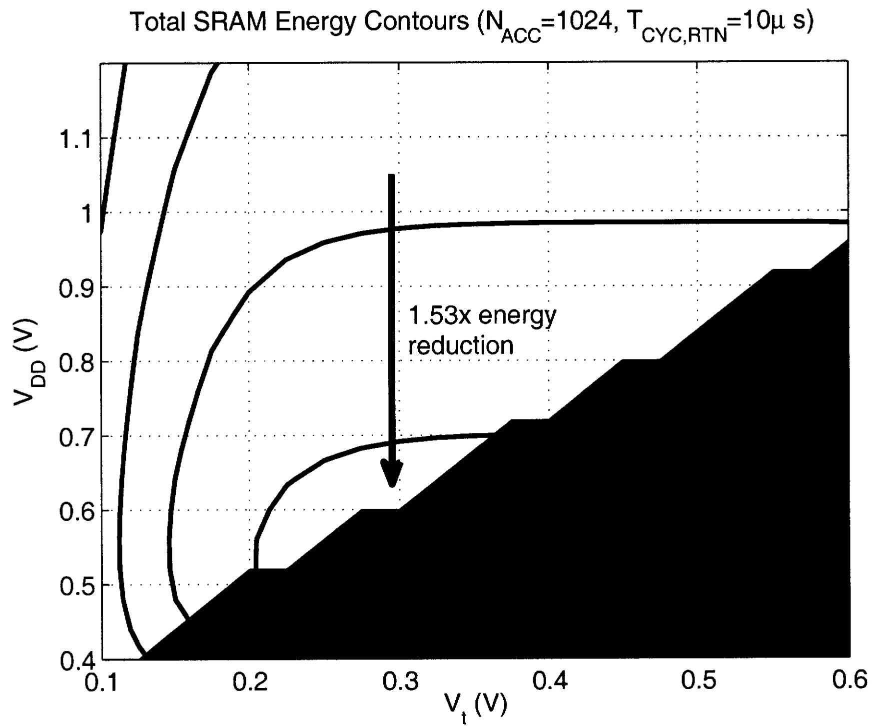
(a) Total energy for $T_{CYC,RTN} = 10ms$.



(b) Total energy for $T_{CYC,RTN} = 1ms$.



(c) Total energy for $T_{CYC,RTN} = 100\mu s$.



(d) Total energy for $T_{CYC,RTN} = 10\mu s$.

Figure 2-19: Sub-array total energy (at room temperature, with variation) for various performance requirements (specified by $T_{CYC,RTN}$).

curs well within the region where the performance constraint is met (i.e. T_{ACC} is considerably less than $10ms$). The reason for this is that long data-retention periods require high V_t in order to minimize leakage-power; however, the severe variation that accompanies this results in extreme performance degradation. The resulting increase in active-mode leakage-energy (E_{LKG}) cannot be offset by the leakage-current savings (which themselves are diminished by the variation gain factor). Consequently, to reduce E_{LKG}, T_{ACC} must be aggressively shortened by raising the supply-voltage. Of course, the net increase in the leakage-energy, and now also the active-energy, is extremely detrimental to the absolute total energy (as quantified below).

The effect on V_{DRV} also has a strong impact on the total energy, particularly in the low performance cases where T_{ACC} must be reduced far below $T_{CYC,RTN}$. For instance, in the $T_{CYC,RTN} = 10ms$ case, the idle-mode power at the optimal V_t (i.e. 0.475V) is higher by over 1.9x compared to the nominal case. This is due to both the higher V_{DRV} and the leakage-current gain factor from variation.

Overall, the increase in T_{ACC}, V_{DRV}, and the aggregate leakage-current has a severe impact on the absolute total energy. Specifically, for the $T_{CYC,RTN}$=$10ms$, $1ms$, $100\mu s$, $10\mu s$ cases respectively, the increases in the total-energies are 1.65x, 1.56x, 1.28x, and 1.33x compared to the nominal cases. From this analysis the most important conclusion that can be drawn is that, for ultra-low-energy applications with modest data-retention periods, it is critical to improve performance in the presence of variation in order to (1) enable aggressive V_{DD} and V_t scaling, and (2) minimize T_{ACC} so that so that the active-mode leakage-energy can be reduced at a given V_{DD} and V_t. As the data-retention period becomes extremely long, it also becomes critical to minimize the idle-mode leakage-power.

2.4 Summary and Conclusions

This chapter investigates the sources of SRAM energy, and, in particular, how they scale with respect to supply-voltage and threshold voltage. Targets and trends for optimal V_{DD} and V_t are established. The analysis here is different from that of generic

digital logic [44] in two important ways: (1) the general requirement of long-term data-retention introduces the need for a low-power idle-mode where the power cannot be assumed to be zero, and (2) due to its extreme impact on SRAMs, variability (and its dependence on V_{DD} and V_t) must be considered.

With the need to persistently retain data, total SRAM energy optimization, based on the analysis of a practical sub-array, points to three necessary trends: (1) lowering V_{DD}, despite the prominence of active-mode leakage-energy, far below the value expected based on generic logic considerations; (2) raising V_t to mitigate leakage-currents during both active- and idle-modes; and (3) reducing the access-delay to reduce the active-mode leakage-energy and increase the amount of V_{DD} and V_t scaling achievable under the performance constraint. Although lowering V_{DD} tends to raise leakage-energy due to the ensuing increase in the access-delay, the inability to completely negate leakage-currents following the active-accesses increases its appeal from the standpoint of active-energy reduction. Nonetheless, analysis of the critical SRAM operating metrics shows that reliable operation at the low target supply-voltages and high target threshold-voltages is vehemently opposed by device variability. In particular, the read-margin (quantified by the read SNM) and the write-margin suffer the primary operational violations. The hold-margin has other important implications, particularly with respect to the minimum data-retention voltage achievable, which degrades as V_t is increased; however, with regards to limiting functionality, the hold-margin is far superseded by read-margin.

The cell read-current is also critical to the sub-array energy, and suffers intolerably at low V_{DD} and high V_t. Specifically, it limits sub-array speed, therefore restricting V_{DD} and V_t scaling in the presence of a performance constraint. Importantly, however, because it is so severely degraded by variation at low supply-voltages and high threshold-voltages, read-current drastically increases the access-delay, elevating the leakage-energy. Consequently, due to the degradation of read-current, enhancing sub-array performance is critical. Unfortunately, both read-margin and read-current must be improved simultaneously, and, as discussed in Chapter 4, design strategies to improve one tend to worsen the other. So, it is beneficial to find some means to improve

performance other than targeting read-current (and without increasing V_{DD} or lowering V_t), in order to reduce leakage-energy and easily overcome the performance constraint, thereby facilitating supply- and threshold-voltage scaling.

The following chapters are guided by the energy optimization targets established in this chapter. Chapter 3 investigates techniques to ensure that the required operating margins and metrics are met in order to operate SRAMs at low supply-voltages and high-threshold-voltages. Chapter 4 investigates techniques to improve performance (while maintaining high density), so that V_{DD} and V_t scaling can be aggressively employed for maximum energy savings.

Chapter 3

Ultra-Low-Voltage SRAM Design

The analysis in Chapter 2 strongly points to voltage scaling in order to minimize the energy of SRAMs. Achieving reliable functionality at low supply-voltages, however, is extremely challenging. In fact, in digital systems, SRAMs pose the primary limitation to low-voltage operation, which is critical for energy-constrained applications. Figure 3-1 shows the minimum supply-voltage achieved by specifically ultra-low-voltage designs that have been recently reported [84]. As shown, logic circuits achieve much better voltage scalability than SRAMs, and the gap is increasing as technology scales. As highlighted in Chapter 2 the reason for this is two-fold: SRAM noise-margins and performance are more sensitive to variation than those of logic circuits, and SRAMs are subject to more extreme levels of variation.

This result suggests that one way to alleviate the low-voltage challenges associated with SRAMs is to target larger geometry technologies, where the effects of variation and leakage-currents are much less severe. Unfortunately, the critical SRAM metric of density is equally important, or perhaps even more important, in highly-energy constrained systems. There are two important classes of energy-constrained applications that specifically benefit from technology and/or density scaling:

(1) **Dynamic Performance Scalable.** Applications such as cellular multimedia handsets [85] and wireless sensor nodes [25] have relaxed workloads for the vast majority of the time, but can require bursts of high performance. Dynamic

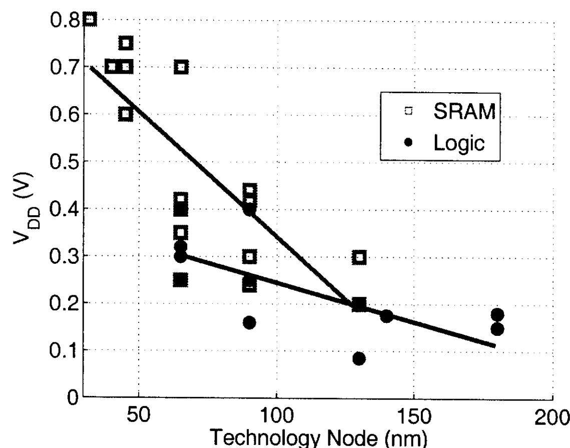


Figure 3-1: Minimum supply-voltage of specifically ultra-low-voltage designs recently reported [84].

voltage scaling and ultra-dynamic voltage scaling [86][87] allows momentary operation at raised voltages, and advanced technologies afford the necessary speed-ups. An important challenge is managing the trade-off between low-voltage functionality and the accompanying overheads to performance and area, which limit high-voltage high-performance operation [88].

(2) **Fixed High-Performance.** Applications such as baseband radio processors [89] and high-resolution video decoders [23] must meet aggressive system throughput specifications. These leverage extreme parallelism, enabled by area scaling in advanced technologies, to operate efficiently at a reduced voltage and frequency [4], which is critical for battery-powered mobile devices.

Despite the challenges and density constraints, the energy optimization analysis in Chapter 2 suggests that SRAM voltage should be reduced to 0.5-0.6V, or even lower (i.e. $< 0.4V$) if the performance degradations imposed by variation can be sufficiently overcome. Accordingly, this chapter describes the design and analysis leading to a 256kb SRAM prototype in LP 65nm CMOS that is intended to operate below 0.4V and achieves operation down to 0.35V. Operation below 0.4V is essential, particularly to ensure engineering margin (of 0.1-0.2V) at the target optimal supply-voltage. Additionally, although the threshold voltage of the implementation

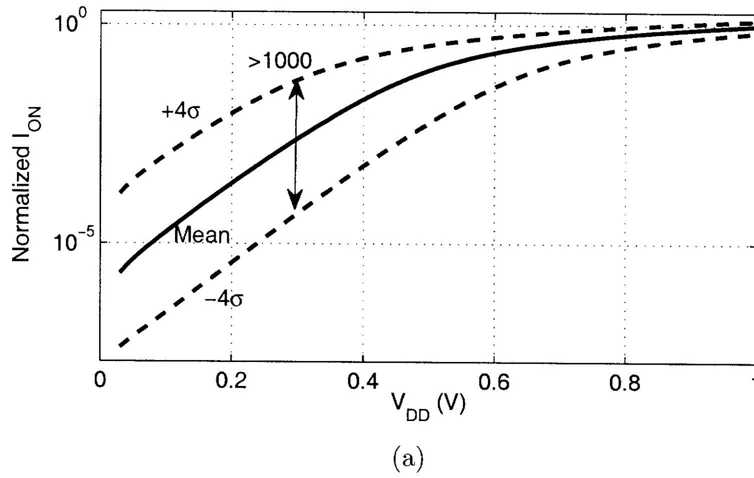
technology is fixed, the analysis of Chapter 2 recommends a higher V_t (up to 0.5V), and a 0.4V demonstration leaves margin for additional device V_t engineering. Finally, Figure 3-1 illustrates the increasing challenge with technology scaling; hence, aggressive low-voltage techniques provide some basis for low-voltage SRAMs in future deeply scaled technologies.

3.1 Low-Voltage SRAM Challenges

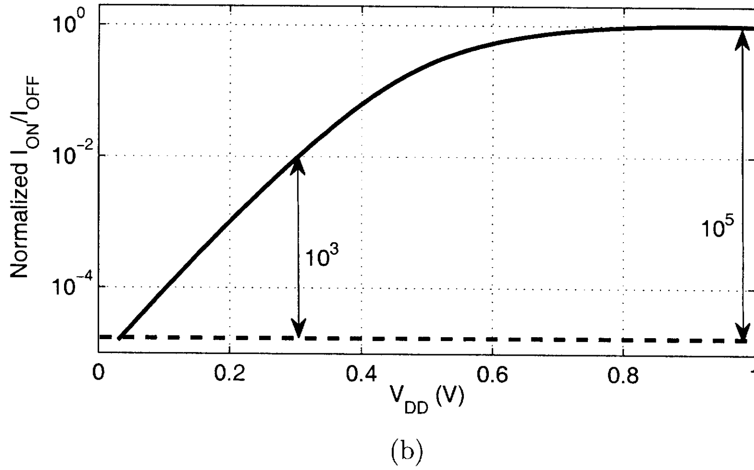
Figure 3-2 shows the normalized I_D versus V_{GS} behavior of an 65nm LP NMOS (predictive model) in two different lights. Specifically, two effects are shown that are of critical importance to SRAMs at low-voltages: 1) the severe effect of threshold-voltage variation (shown in Figure 3-2a), and 2) the degradation in the on-to-off ratio of the drain-current (shown in Figure 3-2b).

In Figure 3-2a, the on-current initially increases exponentially in sub-threshold, and then far more slowly in strong-inversion. As mentioned in Chapter 2, threshold voltage shifts are a prominent result of processing variation and RDF [90][91], and they essentially cause sideways offsets. The $\pm 4\sigma$ case for local-variation, which occurs commonly in large SRAM arrays, is shown. Although the resulting variability is relatively small at high voltages, the change in drain-current is severe at ultra low-voltages (e.g. 0.3V) and can easily exceed three orders of magnitude. This suggests that relative device strengths cannot reliably be set using conventional techniques like W/L sizing. As described in Chapter 2, standard high-density SRAM topologies rely heavily on ratiometric sizing, making them extremely sensitive to this failure mechanism.

Figure 3-2b plots the ratio of the on-current to the off-current for the same 65nm LP NMOS. As shown, the nominal ratio of I_{ON}/I_{OFF} degrades from over 10^5 at nominal voltages to less than 10^3 at 0.3V. The impact of this effect is even more severe when the variation picture from above is considered; for instance, with 4σ degradation to I_{ON}, the ratio at low-voltages is reduced to less than 10^2. This implies that there is now a strong interaction between both the "on" and the "off" devices when it comes



(a)



(b)

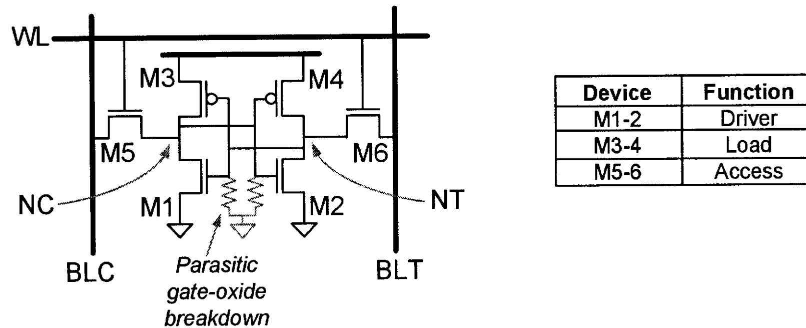
Figure 3-2: Degradation of LP 65nm NMOS (predicitive model) with respect to V_{DD} showing (a) drain-current variation and (b) I_{ON}/I_{OFF}.

to setting the static voltages of critical nodes. This, once again, is highly problematic for SRAMs, where high-density requirements call for the integration of many devices on shared nodes, such as bit-lines.

In the following sub-sections, the basic degradations to MOSFET operation that are brought on at low-voltages are related more specifically to the design of high-density SRAMs. First, the precise effect on bit-cells is considered, specifically with respect to the standard 6T topology. Bit-cell modifications that have been used to address the deficiencies and afford some degree of voltage scalability are also discussed. Then, the challenges related to low-voltage periphery are considered.

3.1.1 Low-Voltage Bit-Cell Array

The 6T bit-cell, which is shown in Figure 3-3, fails to operate at ultra low-voltages because of reduced signal levels and increased sensitivity to threshold-voltage variation [79]. With this topology, both read and write accesses are ratioed making it very difficult to overcome the severe effects of variation and manufacturing defects.



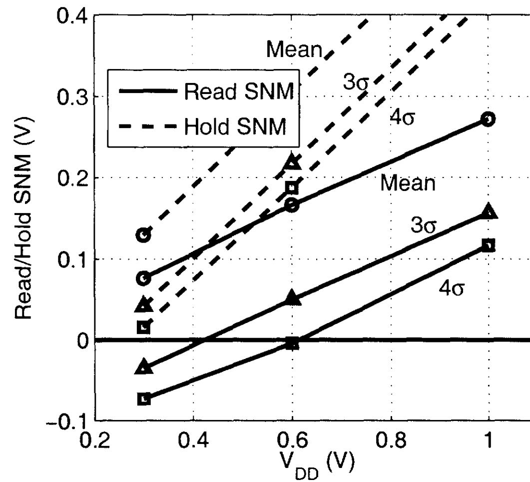
Device	Function
M1-2	Driver
M3-4	Load
M5-6	Access

Figure 3-3: 6T bit-cell for low-voltage analysis.

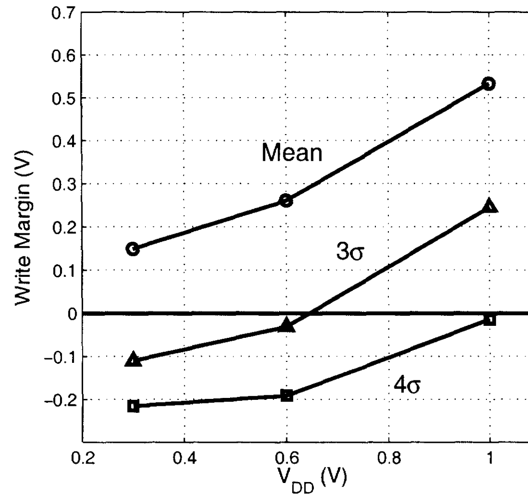
Read/Write Margin

During read-accesses the cell must remain bi-stable to ensure that either data logic value that might be stored can be held and read without being upset by transients that occur at the internal nodes, NC/NT, when the access-devices, $M5-6$, are enabled. The read SNM is considered in Chapter 2 with respect to V_{DD} and V_t, and it quantifies the margin against loss of bi-stability by considering the worst-case condition where the bit-lines (BLT/BLC) remain at the pre-charge voltage of V_{DD}. It should be noted, that in low-voltage designs, the cell read-current is likely to be very small (in proportion to the bit-line capacitance), and BLT/BLC do in fact remain close to the pre-charge voltage for an extended period.

Figure 3-4a shows Monte Carlo simulations of a 65nm LP CMOS bit-cell designed to fit in a layout area of $0.5\mu m^2$ while meeting SRAM design rules for the technology. As expected from the analysis in Chapter 2, at low-voltages the read SNM is violated. Similarly, Figure 3-4b shows how the write-margin is lost at low voltages, which indicates that the cell cannot be made mono-stable [81] at the data state desired to



(a)



(b)

Figure 3-4: $0.5\mu m^2$ 6T bit-cell degradation of (a) read/hold SNM and (b) write-margin with respect to V_{DD}.

be written. Finally, the hold SNM (Figure 3-4a) measures the ability of the cell to remain bi-stable while the access-devices are disabled. As shown, it is preserved to very low voltages and forms the basis for several bit-cell topologies that modify the 6T structure in order to operate at ultra-low-voltages.

Generally, the read and write failures discussed above arise due to the reduced signal levels at low voltages and due to threshold-voltage variation, whose impact also increases at low voltages. The electrical-β ratio [92] is shown in Figure 3-5, and it is defined as the ratio of the effective strength of the driver-devices, $M1-2$, to the

access-devices, $M5-6$. As a result, it is a critical metric characterizing the cell's immunity against problematic transients on NT/NC during read-accesses. In fact, it serves to isolate the contribution that variations in the driver and access-devices have towards read failures. Figure 3-5 shows that the electrical-β ratio, which is nominally set between 2-3 [93], can degrade by almost four orders of magnitude at ultra-low-voltages.

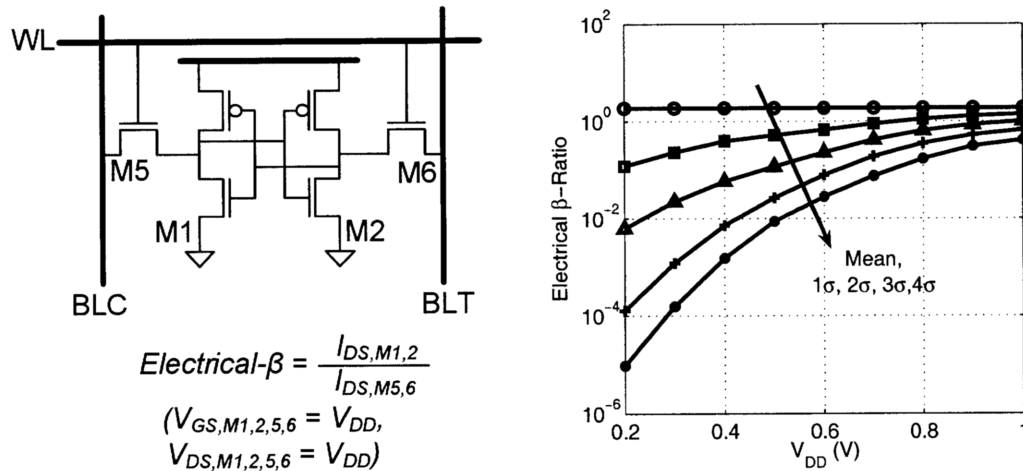


Figure 3-5: Electrical-β ratio definition and degradation with respect to V_{DD}.

In addition to variation, an equally important effect limiting the minimum supply voltage of a 6T bit-cell is gate-oxide soft-breakdown, resulting in extremely high gate-leakage from the driver-devices, $M1-2$ [94]. In 65nm and beyond, even with very high-quality oxide, soft-breakdown unfavorably distorts the read butterfly curves. This can be envisioned by considering the additional parasitic current path shown in Figure 3-3, which at low-voltages can be comparable to the PMOS load-device currents. As a result, this current path lowers the V_{GS} of one of the driver devices, exacerbating the degradation to the logic "0" level of NT/NC during read-accesses [95]. In this manner, soft-oxide breakdown severely limits the minimum voltage where the read-margin is met.

Bit-Line Leakage

In addition to the performance and energy limitations discussed in Chapter 2, the rapid degradation of cell read-current that accompanies V_{DD} scaling also introduces critical functionality failures. Specifically, Figure 3-6 shows the worst-case read-data sensing scenario on one pair of bit-lines. Typically, when a bit-cell is accessed, a droop can be detected differentially on one of the bit-lines, BLT/BLC, with respect to the other. This requires that the read-current of the accessed-cell discharges the intended bit-line more rapidly than the aggregate leakage-current, $I_{LKG,BL}$, which is imposed on the alternate bit-line. Importantly, however, the aggregate leakage-current on the alternate bit-line depends on the data stored in all of the involved unaccessed bit-cells. In the case shown, sense-ability is most severely limited; the stored data in the unaccessed cells leads to high leakage-current on the alternate bit-line (due to the high V_{DS} across the associated access-devices), and nearly no leakage-current contribution to the intended read-current.

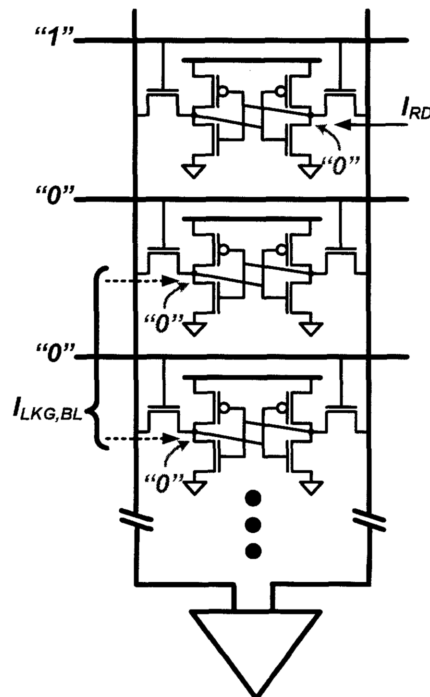


Figure 3-6: Bit-line leakage during read-data sensing opposing the ability to detect differential droops.

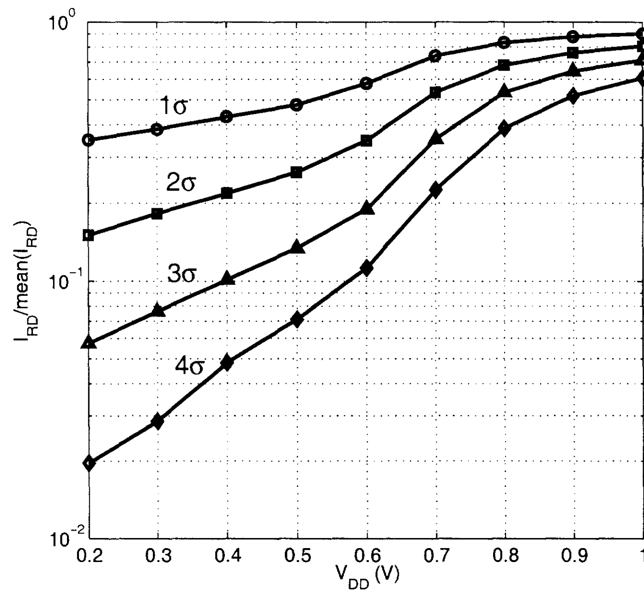
Accordingly, the read-current must be much greater than the worst-case aggregate leakage-current. As mentioned in Chapter 2, however, V_{DD} and V_t scaling (for optimal energy) greatly reduces the read-current, especially in the presence of variation. The severe effect due to variation is shown by plotting the weakened-cell read-currents normalized to the already reduced mean read-current. This is shown in Figure 3-7a for a $0.5\mu m^2$ cell in the target 65nm LP technology. Combining this effect with the I_{ON}/I_{OFF} degradation discussed previously, Figure 3-7b plots the worst-case I_{RD} as a ratio of the worst-case bit-line leakage-current (assuming 256 cells per bit-line). As shown, at low-voltages, the worst-case I_{RD} is exceeded by the worst-case $I_{LKG,BL}$, making data sensing impossible.

The design trade-offs between electrical-β ratio and the cell read-current suggest that the symmetric-6T topology imposes inherent restrictions to ultra-low-voltage operation. The electrical-β ratio can be increased by reducing the strength of the access-devices; however, this degrades the cell read-current, fatally limiting the ability to sense read-data. Alternatively, the electrical-β ratio can be increased by up-sizing the driver-devices ($M1-2$). However, the up-sizing required to overcome the degradation shown in Figure 3-5 is too drastic considering the cost on density. Additionally, a large increase in gate area for the driver-devices can exacerbate the limiting effect of gate-oxide soft-breakdown [95], somewhat opposing the read-margin improvement. To overcome these limitations, alternate bit-cell topologies have been proposed that attempt to improve the low-voltage trade-offs and provide some additional voltage scalability. These are described below as non-buffered-read and buffered-read bit-cells.

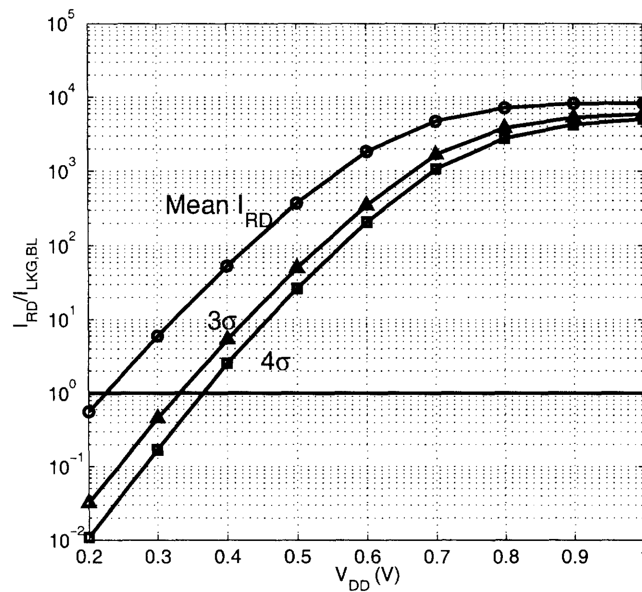
Non-Buffered-Read Bit-Cells

The key benefit of the 6T bit-cell, compared with other static storage structures, is its possibility for maximum density. Whether this possibility can be practically realized, however, is subject to several factors including variability (at the target V_{DD} and V_t), array configuration, and selective biasing conditions during read/write operations.

As mentioned, the difficulty arises due to the need to simultaneously increase



(a)



(b)

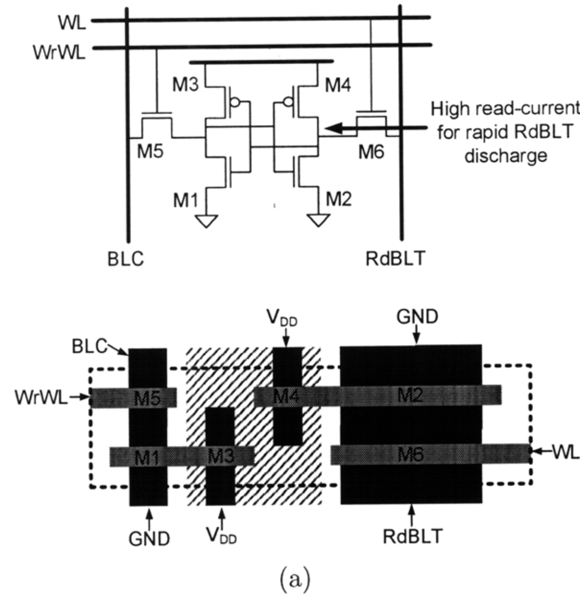
Figure 3-7: Read-current degradation in the presence of variation (a) with respect to V_{DD} scaling and (b) leading to loss of data sense-ability due to bit-line leakage.

read-margin and read-current while lowering the voltage. For instance, several approaches exist to overcome variation and enforce the required read-margin; the electrical strength of the access-devices and load-devices can be altered selectively for read and write operations through precise and adaptive biasing of the word-lines,

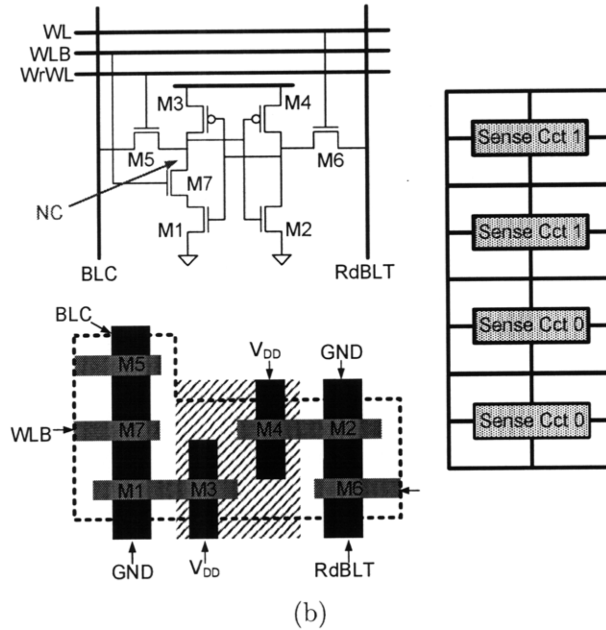
lending to variation resilient cell layout [96] and global process-skew correction [39]. Unfortunately, however, this degrades the read-current [39] to an intolerable level at low-voltages. Alternatively, raising the supply voltage for only the bit-cells being accessed [53] improves the read-margin and read-current but adversely affects the active-access-energy; although the leakage-energy is reduced, the energy saving trends in Chapter 2 cannot be achieved.

Minimally, the read-margin and read-current trade-off at low-voltages can be alleviated by maximizing the discharge rate of the bit-lines, reducing the disruptive stress on the bit-cell storage nodes (NT/NC) that tends to raise their voltage. This is the basis for regenerating the bit-line voltages using a sense-amplifier, as in [68]. Alternatively, the bit-line loading can be aggressively reduced (to less than 16), and the bit-cell pull-down stack on one side can be aggressively up-sized, leading to the asymmetric cell shown in Figure 3-8a [97]. Though the read SNM condition is somewhat eased, severe variation at low-voltages makes it difficult to guarantee that the read-stress is sufficiently eliminated; further, array area-efficiency must suffer, arising from the need to drastically shorten the bit-lines. To completely overcome the read SNM condition, the 7T cell in Figure 3-8b [98] can be used, where the bit-cell feedback path is gated by $M7$ during read-accesses so that the possibility of actively flipping the stored data-state is eliminated. The additional of $M7$ leads to an L-shaped layout, leaving gaps for sensing circuitry to be distributed throughout the columns so that $RdBLT$ can be kept short in order to ensure minimal read access-delay. However, at ultra-low-voltages, the read access-delay is still quite long. Since NC is held dynamically, leakage-currents severely compromise data-storage during these periods, limiting the level of voltage scaling that can reliably be achieved with this bit-cell.

The write operations are somewhat less constrained than read operations, which depend on both the read-margin (to avoid data-disruption) and the read-current (to ensure data sense-ability and achieve optimal sub-array energy). The critical 6T SRAM metric associated with write operations is the write-margin, and it is possible to apply selective biasing to improve the write-margin without simultaneously degrading any other critical metric. In particular, correct write operation requires



(a)



(b)

Figure 3-8: Non-buffered bit-cells formed by (a) asymmetrically upsizing one pull-down path for rapid *RdBLT* discharge [97], and (2) addition of device ($M7$) to gate bit-cell feedback path against disruption [98].

that the NMOS access-devices overpower the PMOS load-devices, in order to store a new data state. A couple of options exist in order to enhance the strength of the access-devices relative to the load-devices. First, the bit-lines that are engaged in the write-operation (recall, column-interleaving implies that only some of the bit-cells in the accessed row must be written to) can be slightly boosted beyond the rail-voltages to over-drive the target access-devices [42]. Second, the supply-voltage

can be reduced, selectively, for the columns that are engaged in the write-operation, thereby weakening their PMOS loads [99]. It is important to note here that the supply-voltage is reduced to all cells in the target columns, including those in the unaccessed rows; however, since their access-devices are disabled they only face the hold-condition, which is far less stringent than the read-condition.

Buffered-Read Bit-Cells

In order to completely eliminate the possibility of read-data-disruption without introducing additional dynamic nodes (which are highly problematic at low-voltages), buffered-read bit-cells can be used. Here, the bit-cell storage-cell, where data is written and held, is explicitly separated from its read-port. This can be done with the addition of at least two devices, leading to the 8T topology shown in Figure 3-9. The evolution from a 6T topology is shown, and the additional devices, $M7-8$, form a read-buffer which isolates the storage-nodes ($M1-6$) from the read-bit-line, $RdBL$. A comparison of the layout for the two cells is also shown, where both adhere to the "Thin Cell" structure [100], which alleviates lithography stresses and device mismatch sources by minimizing jogs in the poly. The 6T layout is limited by the pitch of four devices, while the 8T layout [93] is limited by the pitch of five devices. This layout is highly efficient in its sharing of source-drain junctions and poly wires with abutting cells.

Nonetheless, the additional area overhead introduced, compared with the 6T bit-cell is unavoidable. Hence, to evaluate the merits of the 8T topology for low-voltage operation, it must be compared against the approach of up-sizing the devices of a 6T topology for variation reduction, which would reduced its failure probability due to read-data-disruption. Further, 6T up-sizing also leads to stronger bit-line pull-down paths, increasing the cell read-current. Consequently, for comparison, the 8T cell must use read-buffer devices ($M7-8$) that are sized equivalently to the 6T driver and access-devices. Accordingly, in the iso-read-current 8T cell, very little area is left for the storage-cell ($M1-6$). Importantly, however, the storage-cell must only meet the hold SNM rather than the far more stringent read SNM, which must be met by

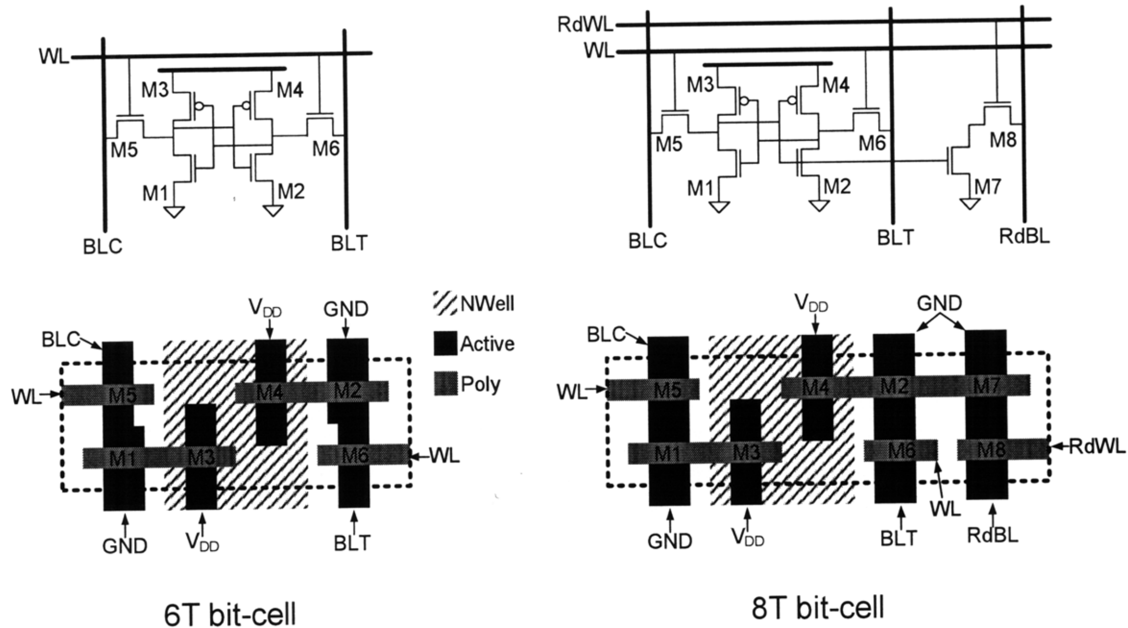


Figure 3-9: 8T bit-cell and layout (to overcome read-data-disruptions) shown besides a typical 6T bit-cell and layout.

the 6T bit-cell. Figure 3-10 compares the 4σ read SNM of 6T bit-cells and the 4σ hold SNM of the storage-element of 8T bit-cells that have been sized for equivalent read-current and total layout area (ranging from $0.65\mu m^2$ to $1.15\mu m^2$) using SRAM design rules in the target 65nm LP technology. As shown, for each given area, the 8T cell achieves lower voltage operation (by approximately 0.2V), as indicated by the V_{DD} where the hold/read SNMs equal zero, respectively.

Since it breaks the highly constrained trade-off between read-margin and read-current, and because its remaining operating margin (i.e. hold SNM) is much more robust to variation, the 8T bit-cell is extremely promising for low-voltage operation. Additionally, for a given layout area, many more options exist to address the low-voltage challenges. For instance, the read-buffer's access-device ($M8$) can be made at least as strong as its driver-device ($M7$), and/or its word-line ($RdWL$) can be boosted, both of which lead to greatly enhanced read-current. Additionally, the threshold-voltages of the storage-devices ($M1-6$) can be raised, to manage leakage-currents (which critically limit the energy savings in both active- and idle-modes), and the threshold-voltages of the read-buffer-devices ($M7-8$) can be lowered to further

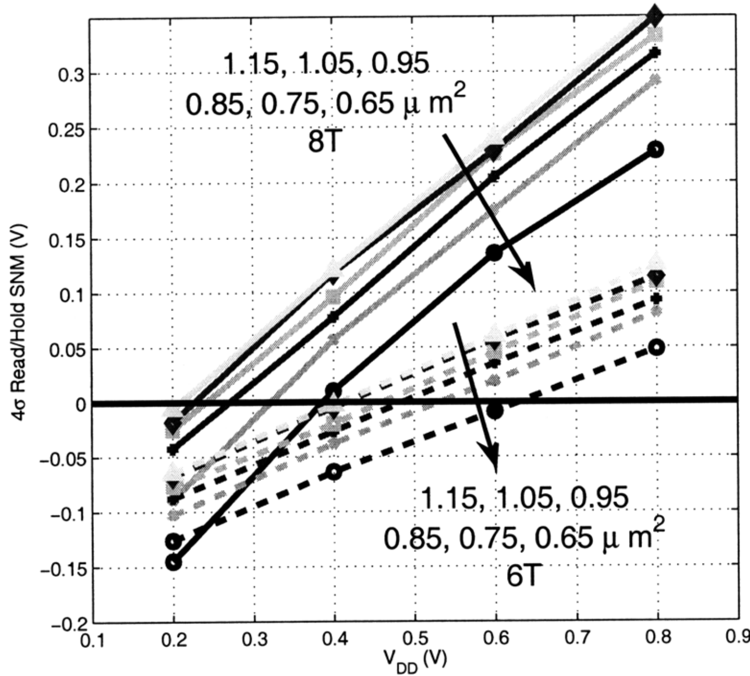


Figure 3-10: 6T bit-cell and 8T bit-cell operating margins for various size layouts (and equivalent read-current) in LP 65nm CMOS.

improve read-current. Finally, with the relaxed operating margin afforded by the hold-condition, the storage-element can be aggressively optimized for write-margin.

Although the 8T bit-cell addresses the matter of read-margin very effectively, it still suffers from bit-line leakage at ultra-low-voltages. To address this, the read-buffer has been enhanced in the bit-cells of Figure 3-11 [101][102]. In Figure 3-11a, when the cell is unaccessed (i.e. $RdWL = 0$) and $NC = 0$, NCB is actively pulled-up to V_{DD} through $M10$. Hence, sub-threshold leakage from $RdBL$ to NCB is eliminated. alternatively, when $NC = 1$, NCB is set by the relative leakage currents from $M9$ and $M10$, and PMOS/NMOS threshold-voltage skews in the technology used lead to higher PMOS leakage currents, once again causing NCB to settle near V_{DD}. In Figure 3-11b, when the cell is unaccessed, NCB is actively driven to V_{DD} by $M10$, regardless of NC. In both bit-cells, however, the read-buffer enhancement imposes an additional area overhead of two devices (i.e. 10T cells).

To overcome read-bit-line leakage without additional area overhead, the design

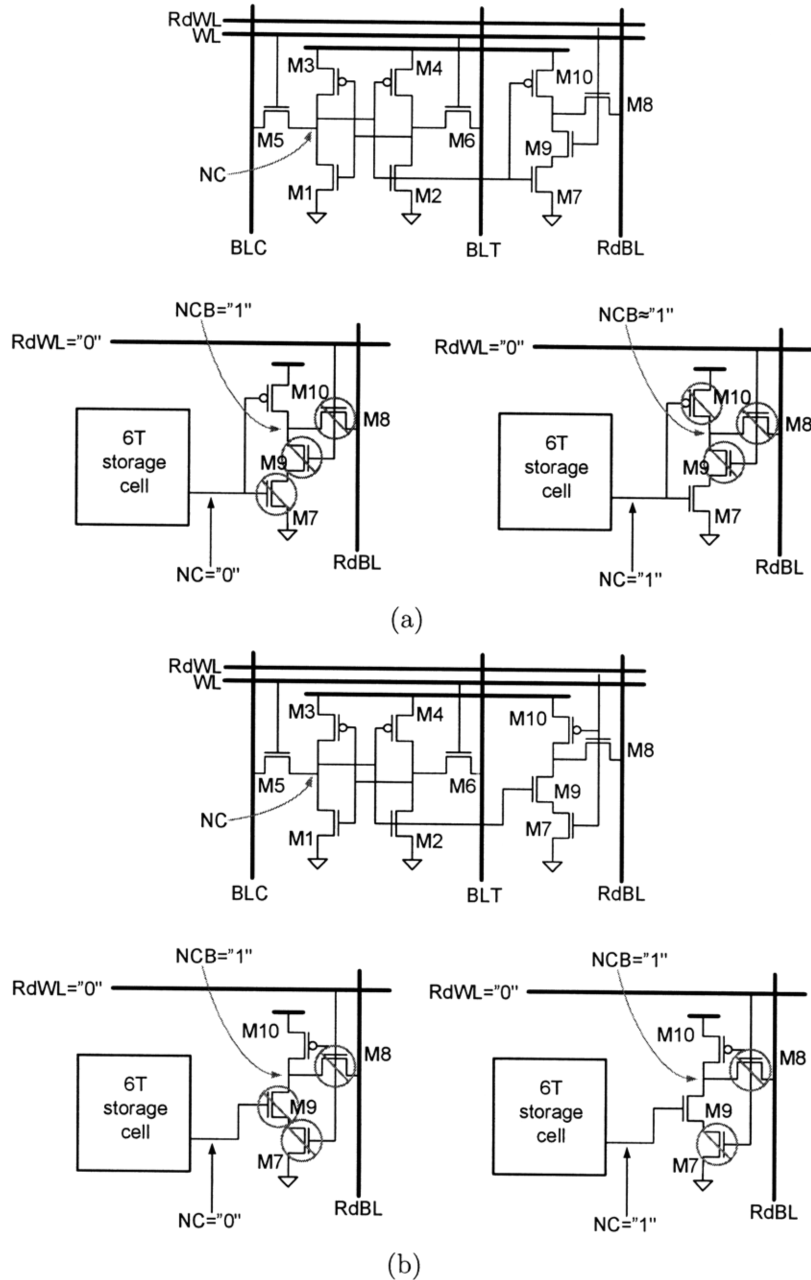


Figure 3-11: Bit-cell read-buffer enhancements to manage bit-line leakage using (a) PMOS/NMOS threshold-voltage skews [101], and (2) active pull-up on internal NCB node [102].

in [103] uses a standard 8T cell in a general-purpose process with lower threshold-voltages. This leads to less variation and higher gate-overdrive on the devices (i.e. improved I_{ON}/I_{OFF}). Consequently, bit-line leakage can be sufficiently managed by shortening the read-bit-lines (i.e. eight cells per read-bit-line are used). The degradation to array area-efficiency is alleviated by using very simple read-bit-line

sensing circuitry and maintaining much longer write-bit-lines (BLT/BLC), which are not susceptible to leakage effects (i.e. 512 cells per write-bit-line). Nonetheless, as discussed in Chapter 2 reducing the threshold-voltage can have a highly detrimental effect on the total energy in low- and medium-performance highly-energy-constrained applications.

An issue with the 8T bit-cell is that it is not as amenable to a column-interleaved layout as the 6T bit-cell. During write-operations in a column-interleaved array, only some cells of the accessed-row are engaged by differentially driving write data on to their bit-lines. The bit-lines of the remaining cells in the row, which are referred to as half-accessed, are precharged high, imposing a read-condition since their access-devices are enabled. Truly eliminating the read-condition, then, restricts column-interleaving, requiring that bit-cells accessed together be layed-out adjacently. Unfortunately, soft-errors, caused by radiation of energetic particles [104], exhibit spatial locality [105], implying that, without column-interleaving, several bits of an accessed word can be corrupted. In order to avoid the increased complexity of multi-bit error correction coding (ECC), soft-error immunity can be achieved by using several ECC bits per row, each corresponding to spatially separated bit-cells. With the 128b per row configuration of [103], the extra ECC penalty is only 5%. Alternatively, a buffered-read bit-cell has also been proposed to enable column-interleaving; however, its area overhead is considerable, requiring a total of ten devices (i.e. 10T) [106].

3.1.2 Low-Voltage Periphery

With respect to overall SRAM density, the area of the periphery, including address decoders, word-line drivers, sense-amplifiers, and read/write data interfaces, is less important than that of the array. As a result, circuit-styles that are much more robust to low-voltage operation can be employed. The peripheral decoders and array control-signal drivers are considered below, followed by the sense-amplifiers

Decoders and Drivers

For decoding circuitry and word-line drivers, static CMOS logic gates, free of ratiometric functionality dependencies and nominally achieving rail-to-rail signal levels, are amenable for low-voltage operation [51][107]. This is particularly true if gates with high input fan-in are avoided, in order to minimize parallel leakage-paths that degrade the output logic level [108]. Further, in the nominal case, sizing of CMOS gates for minimum energy also favors high density [109]. In the presence of variation, however, care in logic gate design is necessary, requiring transistor up-sizing to ensure that actively "on" pull-up/down networks are not so severely weakened that that they cannot compete with their complementary pull-down/up networks (which are nominally "off") [110]. Nonetheless, unlike bit-cells, which exhibit extreme levels of variation (i.e. 4-5σ) in large arrays, the periphery can be designed for much less variation, on the order of 3σ. Consequently, comparatively modest up-sizing is required to ensure low-voltage operation.

Sense-Amplifiers

Unlike decoder and driver circuitry, sense-amplifiers have much more stringent operational requirements than static CMOS logic gates. Even if full-swing bit-line sensing is used [111], the bit-line logic-levels are severely degraded at low-voltages by two critical factors: (1) severe bit-line leakage, arising since the paradigm of low nodal fan-out [108] can not be up-held on the bit-lines, and (2) drastically reduced bit-cell read-currents, arising since large arrays are subject to to extreme levels of bit-cell variation.

Of course, by using very short read-bit-lines (e.g. eight), simple CMOS-style full-swing logic sense-amplifiers can be used [103]. However, in high-density array configurations, where 256 cells are integrated per bit-line, sense-amplifier inverters pose one of the primary limitations at low-voltages [81]. To reduce the sensing margin required, the sense-amplifier inverters' trip-points can be adjusted by using a replica column to emulate the logic levels that have been compromised by bit-line leakage.

This technique is effective in the 130nm design of [102], where the aggregate bit-line leakage is less than the worst-case read-current, and read-current variation is sufficiently small to have minimal impact on the bit-line logic-levels.

In advanced technologies, however, low read-current variation cannot be presumed; further, variation in the sense-amplifiers themselves greatly increases the required sensing margin. To manage the effect of global variation in the sense-amplifiers, differential structures, such as the strong-arm flip-flop (SAFF) [112] or the conventional input-regenerating latch [113], can be used. Differential buffered-read bit-cells, such as the 9T design of [114] and the 10T design of [106], provide compatibility at the cost of density, though pseudo-differential sensing can also be employed, as discussed in Section 3.2.2. Unfortunately, these approaches do not address the critical concern of local-variation in the sense-amplifiers, which poses a primary limitation to their own area scaling, and, increasingly, that of the entire array [111].

3.2 Ultra-Low-Voltage SRAM Prototype

In this section the design, implementation, and testing of a low-voltage SRAM prototype is described. Solutions that are practical with respect to the overheads they introduce are presented to address the challenges discussed in Section 3.1. Importantly, the design presented is meant to be compatible with technology directions specifically targeting low-energy and high-density, even though these aggravate variability. Specifically, a 65nm LP CMOS technology is used. Although the device threshold-voltages are preferentially optimized for low-power, even higher V_t would be desirable for the severely energy-constrained applications discussed in Chapter 1. Consequently, aggressive voltage scaling is pursued in order to leave margin for further V_t optimization as well as practical guard-band from supply/ground fluctuations. The prototype is designed to operate below 0.4V, corresponding to a sub-threshold supply-voltage. Furthermore, to maximize array area-efficiency and specifically target the challenge of bit-line leakage at low-voltages, 256 cells are integrated per bit-line. Finally, the total capacity of the prototype is 256kb, which can serve as a reasonable

size cache for many practical low-energy applications. Accordingly, this design forms the basis for the SRAM used in actual systems that have been prototyped, including a low-voltage DSP [11] and H.264 video decoder [23].

3.2.1 8T Bit-Cell with Low-Voltage Circuit Assists

The prototype uses the bit-cell shown in Figure 3-12. It is based on the 8T topology, employing a 6T storage-cell and a 2T read-buffer which isolates the storage-cell during read-accesses. Consequently, as discussed in Section 3.1, the read SNM limitation is eliminated. The other two prominent limitations, namely bit-line leakage and write-ability in the presence of variation, are dealt with using the peripheral assists associated with the $BffrFt$ and VV_{DD} controls, which are described below.

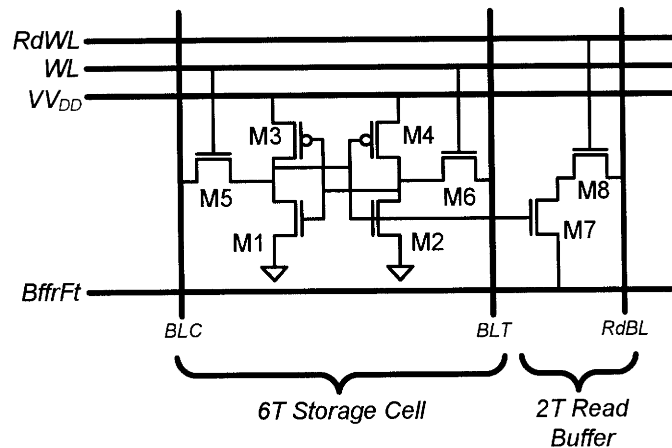


Figure 3-12: 8T bit-cell uses two-port topology to eliminate read SNM and peripheral assists, controlling $BffrFt$ and VV_{DD}, to manage bit-line leakage and write errors.

Bit-Line Leakage Assist

The bit-line leakage problem in the single-ended 8T cell is analogous to the problem in the 6T case (discussed in Section 3.1), except that the leakage-currents from the unaccessed cells and the read-current from the accessed cell affect the same node, $RdBL$. Consequently, the parasitic leakage-currents can pull down $RdBL$ regardless of the accessed cell's state. Figure 3-13a shows transient simulations where $RdBL$ is

correctly pulled low by the accessed cell in the solid curve, but it is also erroneously pulled low in the dotted curves by the leakage currents of the unaccessed cells. Here, only the case with 64 cells on $RdBL$ results in any sampling window; of course, the need for additional engineering margin limits the achievable integration even further.

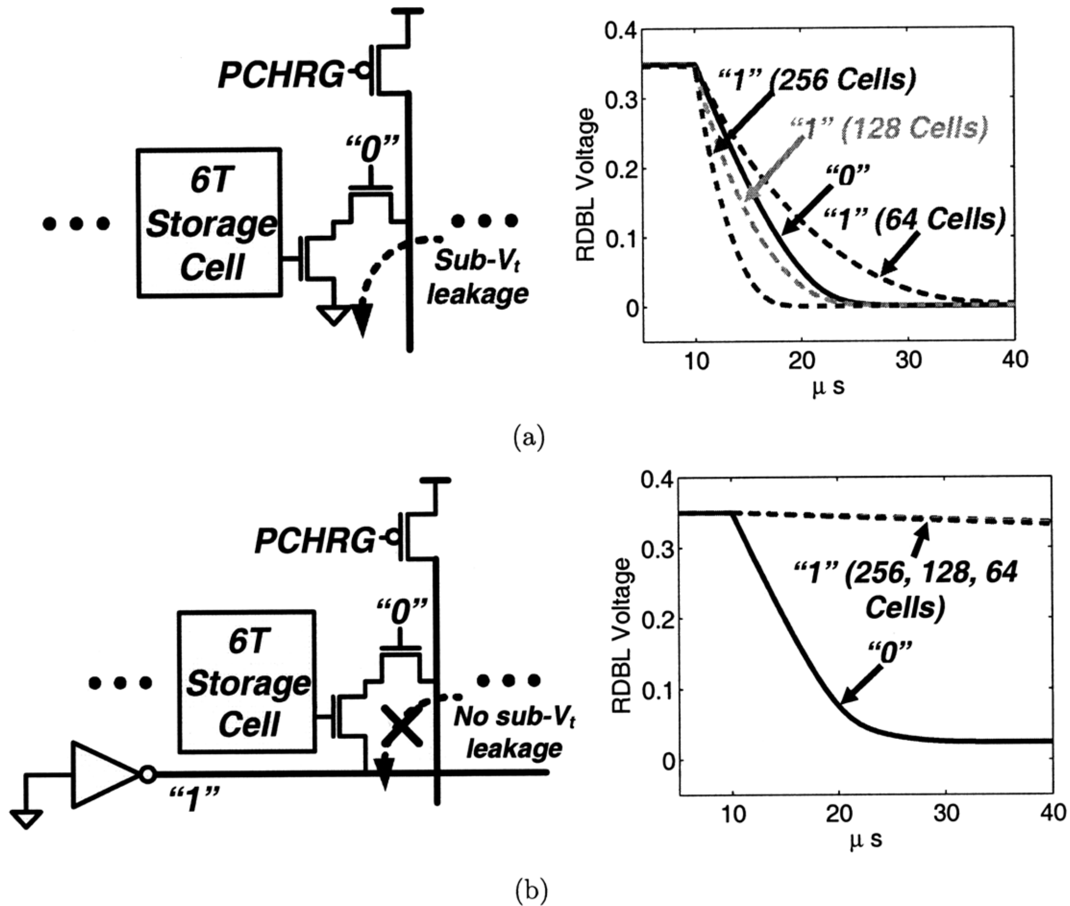


Figure 3-13: Read-buffer bit-line leakage in (a) conventional case where unaccessed read-buffer foot is statically connected to ground and (b) this design where unaccessed read-buffer foot is pulled up to V_{DD}.

In this design, however, the feet of all the unaccessed read-buffers are pulled up to V_{DD} through the $BffrFt$ control, as shown in Figure 3-13b. Consequently, after $RdBL$ is precharged, the read-buffer devices have no voltage drop across them, and they sink no sub-threshold leakage-current. The transient simulation in Figure 3-13b now shows that $RdBL$ correctly remains high in the dotted curves even when 256 cells are integrated. Some residual droop is still visible; this comes about as a result of gate-leakage from the read-buffers' access-devices and junction-leakage from their

drains. Although some energy overhead is incurred to switch $BffrFt$, this is roughly equivalent to WL assertion, and it is much less than the energy of the aggregate BL switch capacitance.

An important concern with this approach is that the peripheral NMOS device of the $BffrFt$ driver must sink the read-current from all cells in the accessed row. As shown in Figure 3-14 (and discussed further in Section 3.2.3), this design has 128 cells per row, making the current requirement of the footer device impractically large. Unfortunately, this device faces a two-sided constraint, and cannot simply be up-sized to meet the required drive strength, since this would introduce excessive leakage-current in the $BffrFt$ driver of the unaccessed rows, and, additionally, the resulting area increase would offset the density advantage of using a peripheral assist.

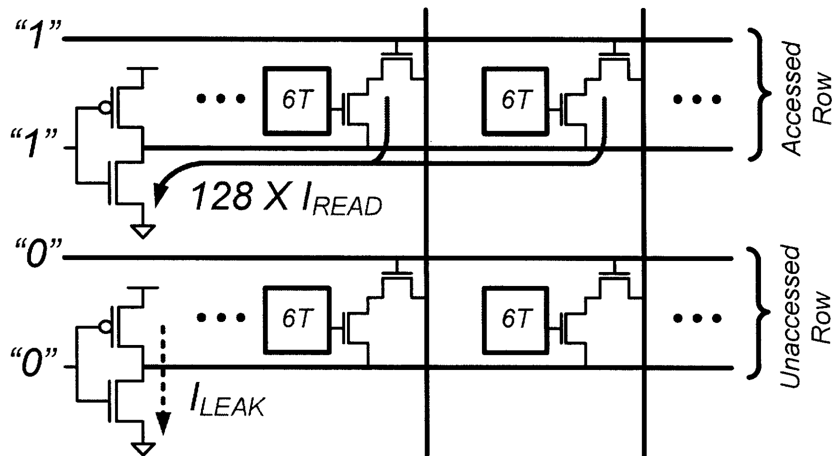


Figure 3-14: $BffrFt$ driver must sink the read-current from all bit-cells in accessed row, and it draws leakage-current in all unaccessed rows.

Instead, the $BffrFt$ driver is itself driven with the charge-pump circuit shown in Figure 3-15a. This ensures that, when accessed, the gate-drive of its NMOS is at least 600mV instead of 350mV, and since this device is in (or very near) sub-threshold, its current increases exponentially, by over a factor of 500, as shown. As a result, the devices of the $BffrFt$ driver can be nearly minimum sized, consuming negligible leakage-power in the unaccessed rows. Additionally, since their gate nodes have minimal capacitance, the charge-pumps and their boost-capacitors (C_{BOOST}) can be physically small, occupying just slightly more area than a couple of bit-cells.

The charge-pump circuit itself is suitable for this ultra-low-voltage application since it uses a PMOS, $M1$, to precharge the boost-capacitor and is free from threshold voltage drops. The transient simulation in Figure 3-15b shows that when a row gets accessed, its BFB node gets bootstrapped to nearly $2V_{DD}$, and the following NMOS can easily pull down the feet of the accessed read-buffers. A side-benefit of this approach is that the read-buffer devices consume no sub-threshold leakage-power themselves. In a typical 8T bit-cell, the read-buffer imposes an additional leakage path. Here, however, that overhead is almost completely mitigated.

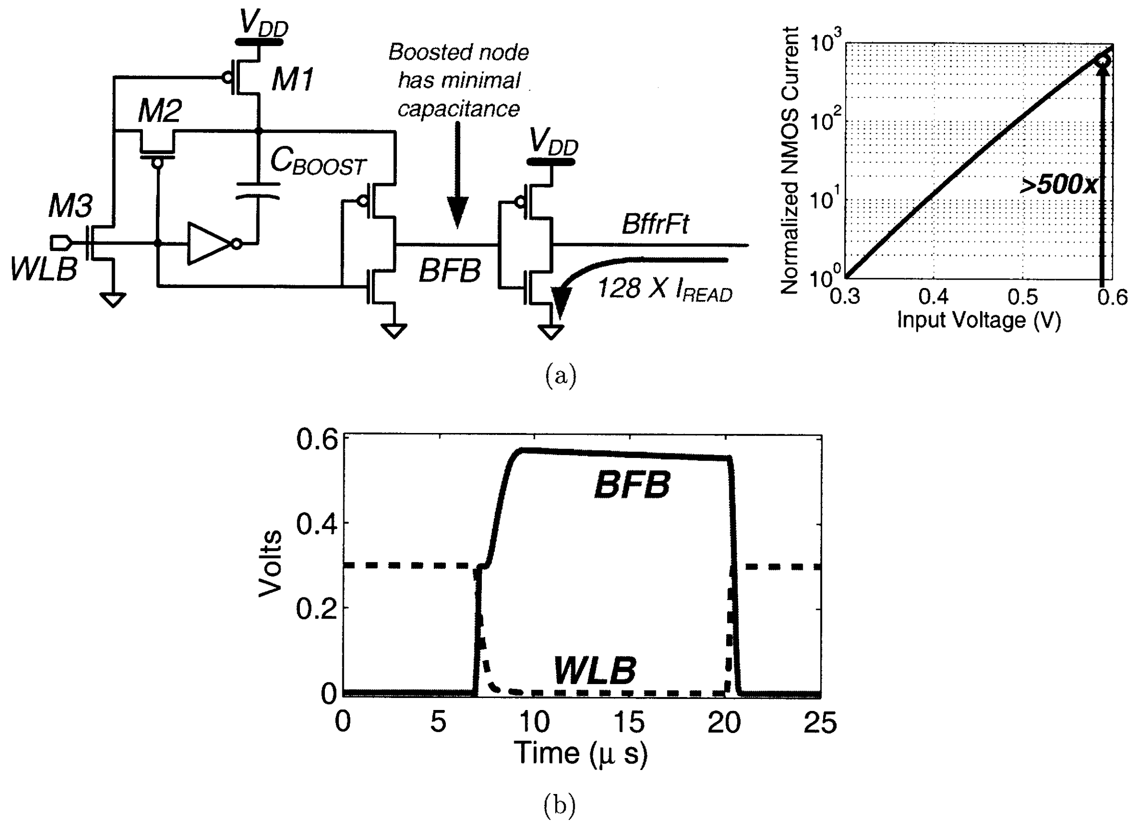


Figure 3-15: To resolve read-buffer footer limitation (a) charge-pump circuit is used (b) BFB node gets bootstrapped to approximately $2V_{DD}$ increasing the current of the $BffrFt$ driver by over 500x.

Write Assist

Write failures occur because, in the presence of variation, it cannot be guaranteed that the strength of the access-devices ($M5-6$) is more than the strength of the load devices ($M3-4$). However, it is possible to enforce the desired relative strengths

using circuit assists. For instance, as mentioned in Section 3.1 the appropriate bit-line voltage can be pulled below ground or, in a non-column-interleaved array, the word-line voltage can be boosted above V_{DD} in order to increase the gate-drive of the NMOS access-devices. Unfortunately, both of these approaches require boosting a large capacitance, either the bit-line or word-line, beyond one of the rails. An alternate strategy that avoids generation of an explicit bias voltage involves weakening the PMOS load devices by reducing the cell supply-voltage. Figure 3-16 shows that as the supply-voltage is reduced, the strength required of the access-devices is eased, which is reflected by the accompanying decrease in the minimum word-line voltage that results in a successful write. So, in this design, write-ability down to 0.35V is ensured by boosting the word-line slightly, by 50mV, but more importantly by reducing the cell supply voltage to weaken the PMOS load devices.

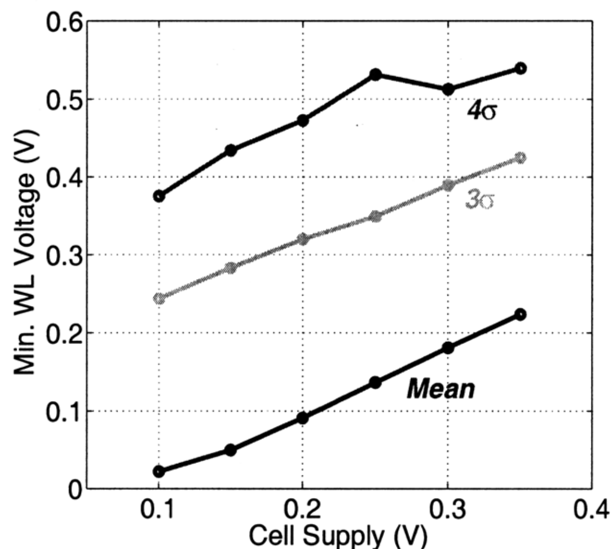


Figure 3-16: Minimum word-line voltage resulting in a successful write with respect to the bit-cell supply voltage.

As shown in Figure 3-17a, all cells in each row share a virtual supply node, labeled VV_{DD}. Previously, VV_{DD} has been reduced passively by disconnecting its power-supply [99][101]. However, to ensure write-ability with more aggressive voltage-scaling in this design, VV_{DD} gets actively pulled low during the first half of the write cycle by a peripheral supply driver. Despite this, as shown in Figure 3-17b, VV_{DD} does

not go all the way to ground because all of the accessed cells contribute to pulling it back up. Specifically, one of the bit-lines gets pulled low, causing the corresponding storage node, NT, to go low. Accordingly, the alternate PMOS load device tends to turn on, introducing a current path from the opposite storage node to VV_{DD}; in this manner half the bit-cell contributes to pulling VV_{DD} back up through one of its PMOS load devices and one of its NMOS access devices. Fortunately, this interaction is quite accurately controllable, since the pull-down devices of the supply driver are large enough that they experience minimal local variation, and the pull-up path through all of the accessed bit-cells tends to stabilize variability. It is important to note, however, that the supply driver does introduce an additional leakage path in all of the unaccessed rows. To minimize its leakage-current, series NMOS pull-down devices are used, taking advantage of the stacked-effect [115].

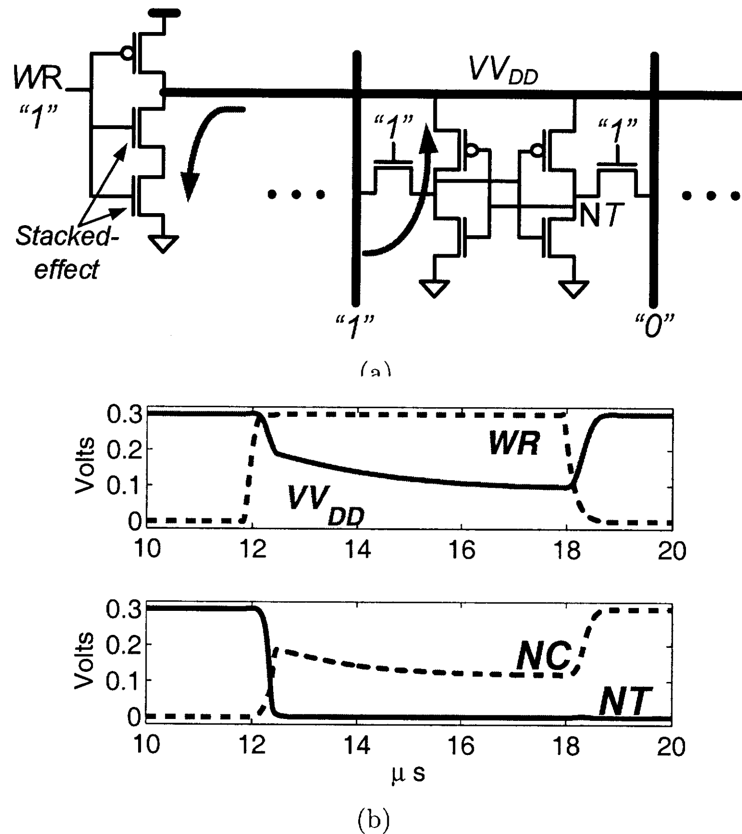
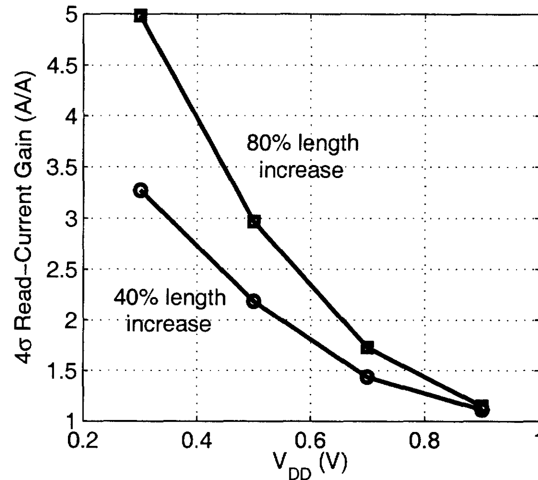


Figure 3-17: Virtual V_{DD} scheme (a) supporting circuits, and (b) simulation waveforms.

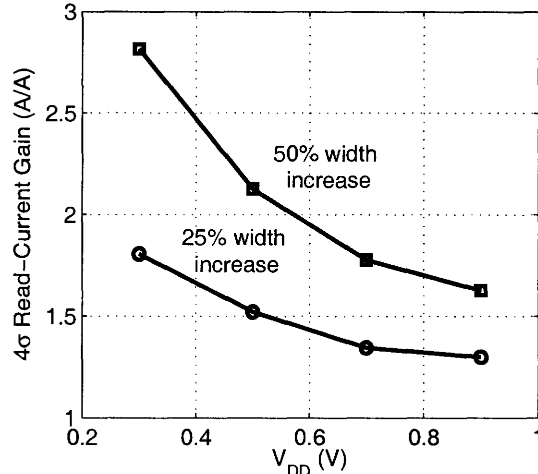
Bit-Cell Layout and Sizing

Figure 3-10 shows that, by eliminating the read-condition on the storage-element ($M1-6$), the 8T bit-cell has greatly eased operating margins, allowing it to be sized much more aggressively. Read-buffer (i.e. $M7-8$) sizing, however, remains an important concern due the effect of read-current on energy, performance, and functionality. Importantly, threshold-voltage variation, whose standard-deviation is inversely related to the square-root of device area, has elevated impact at reduced voltages (approaching an exponential dependence towards sub-threshold); as a result, read-buffer up-sizing has significantly enhanced appeal. Figure 3-18 shows the benefit, at the 4σ level, of upsizing the read-buffer beyond the dimensions used for a $0.65\mu m^2$ cell in the target 65nm LP technology. As shown, even modest increase in the width and length improves the current gain greatly at low-voltages. The length increase is particularly beneficial, since, in addition to alleviating variability, it improves even the mean read-current by effectively reducing the device threshold voltages through the reverse-short-channel effect (RSCE); here, increasing length reduces the influence of high-concentration "halo" doping in the channel region [116]. This effect can also be used to improve the strength of the storage-cell access devices to improve the write-margin [117].

With regards to the conventional 8T bit-cell layout, length increase by approximately 40% of the read-buffer devices can be achieved with nearly no area overhead. As shown in Figure 3-19, the layout height is limited by two minimum length NMOS devices ($M1,5$ and $M2,6$) that require an active-contact between them. However, since the read-buffer devices require no contact to their intermediate node, their lengths can be increased as shown. Additionally, the 8T bit-cell used in this design requires row-wise control of the $BffrFt$ node. This can be achieved with no additional overhead, as shown; in this manner, the $BffrFt$ node is shared by two rows, but the resulting increase in bit-line due to $BffrFt$ activation in the unaccessed cell is negligible. If VV_{DD} control is not required (i.e. the write-margin is sufficiently enhanced by storage-cell sizing and device engineering, as suggested in Section 3.1.1),



(a)



(b)

Figure 3-18: Read-current gain as a result of read-buffer upsizing (a) via width increase, and (b) via length increase (taking advantage of reduced variability and RSCE).

the remainder of the layout can be left unchanged.

In this design, however, VV_{DD} control is required to ensure that the write-margin does not limit voltage scaling. Further, logic design rules are used (instead of SRAM rules), leading to a larger bit-cell layout that must be redesigned to minimize area. The final layout used is shown in Figure 3-20, where all devices have been rotated, isolating the VV_{DD} node. Additionally, to share VV_{DD} routing/contacts, $RdBL$ routing/contacts, and PMOS N-wells, the row is physically folded [118], as shown in the

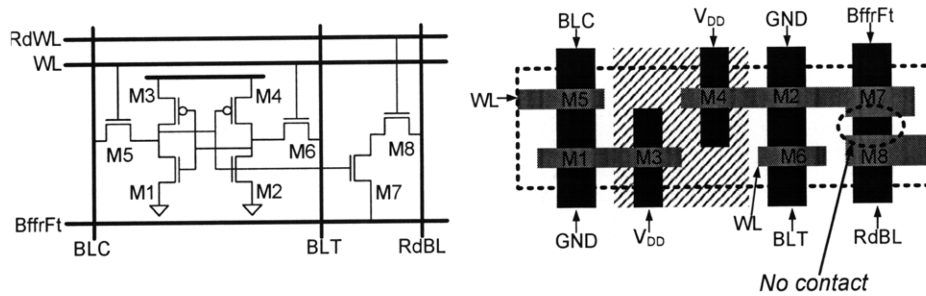


Figure 3-19: 8T bit-cell layout with read-buffer upsizing and $BffrFt$ control (but no VV_{DD} control).

layout tiling.

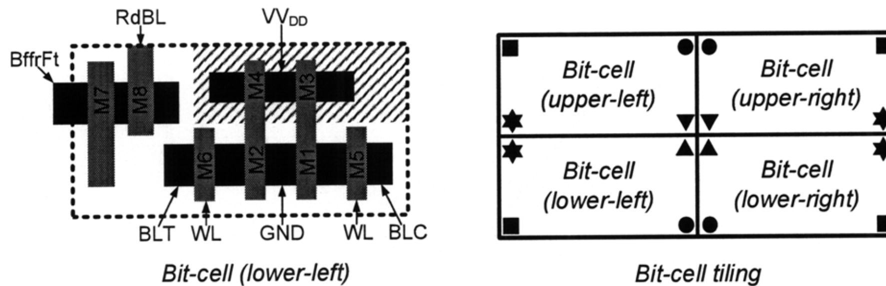


Figure 3-20: Final 8T bit-cell layout and folded-row tiling.

3.2.2 Sense-Amplifier Redundancy

An important consequence of using a non-column-interleaved layout (described further in Section 3.2.3) is that adjacent columns can no longer share a sense-amplifier. As a result, each column must have its own, making the area of each sense-amplifier more constrained and increasing the total number in the entire SRAM. This scenario stresses a general problem observed in deeply scaled technologies. Specifically, the size of the sense-amplifiers has stopped scaling due to the trade-off between their statistical offset and their physical size [111]. In this design, that trade-off is managed in part by using a "full-swing" sensing scheme, where the read-bit-line is allowed to discharge completely. Considering the significant speed-up conventionally obtained by small-signal sensing, this might seem like a drastic approach. However, as mentioned in Section 3.2.1 the unaccessed read-buffers do impose some minimal droop on their

$RdBL$ due to gate and junction leakage. Conversely, as the $RdBL$ voltage level falls, the unaccessed read-buffers start to drive reverse sub-threshold leakage current from their $BffrFt$ nodes, which are at V_{DD}, on to the $RdBL$ node. The resulting droop ultimately settles to approximately 120mV. Unfortunately, as mentioned in Section 3.1, read-current variation can cause the read-access time to extend almost arbitrarily, and, in fact, it approaches the settling time of the transient droop. Consequently, in this design, a static discipline is adopted that guarantees that the correct data value can be sensed on the read-bit-line even after the read-current and droop transients have settled. Specifically, this implies that the offset of all of the sense-amps must be bound by the 120mV logic "1" level and ground logic "0" level of $RdBL$. To achieve this offset under the imposed area constraints, sense-amplifier redundancy is employed, as described in the following sub-sections.

Sense-Amplifier Offset Sources

Offsets in sense-amps come about as a result of global and local-variation in their devices [41]. Here, global-variation refers to die-to-die variation in devices, and local-variation refers to mismatch between devices within the same die placed close to each other. Global-variation can affect all of the NMOS devices on the chip differently than the PMOS devices, thereby, for instance, skewing the switching threshold of all inverters. Alternatively, local-variation can affect the switching threshold differently for each inverter.

Importantly, the effect of global-variation can be cancelled by using a differential sense-amp, as shown in Figure 3-21. The symmetry in this structure ensures that the devices in its two branches will not be subject to systematic differences in process variation. Of course, the 8T bit-cell of this design uses a single-ended read-buffer and is incompatible with differential sensing. Accordingly, pseudo-differential sensing is used, where the actual read-bit-line drives one of the inputs in Figure 3-21, and an off-chip reference drives the other high-impedance input. Although this off-chip reference provides valuable testability support, for a more integrated solution, the PMOS input device on the reference side can be separated into two devices that are

driven by replica columns providing the worst-case $RdBL$ logic "1" and "0" levels. Assuming the reference is properly generated, the differential signal on the read-bit-line that must be resolved is 60mV.

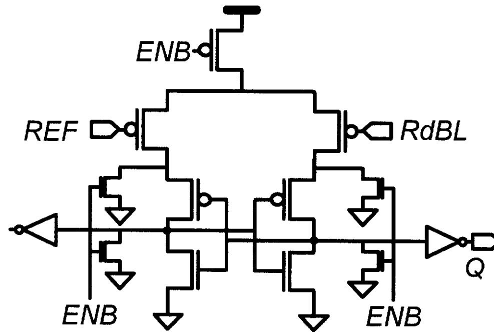


Figure 3-21: Differential sense-amp structure cancels effects of global variation.

The remaining source of offset is local-variation, which is modeled as a random effect whose standard-deviation is inversely related to the square-root of the device areas [119][120]. This gives rise to the area-offset trade-off that is also shown in the Monte Carlo curves of Figure 3-22. In this design, where there are a total of 1024 sense-amplifiers (as discussed in Section 3.2.3), considerable up-sizing would be required to keep the number of failures from offset to an acceptable limit.

Sense-Amplifier Redundancy Concept

As shown in Figure 3-23, sense-amplifier redundancy requires that the read-bit-line ($RdBL$) from each column be connected to N different sense-amplifiers. Each of these has the differential structure shown in Figure 3-21, so their offsets are from local-variation, and they are therefore non-systematic and uncorrelated. Only one sense-amplifier from among the N is selected, whose offset is bound by the high and low logic levels of $RdBL$. So, if the selection can be made correctly, only one of the N sense-amplifiers must have sufficiently low offset. A similar approach has been applied to flash ADCs to achieve minimal offset in the thermometer coded comparators [121].

Importantly, though, the total area for all of the sense-amplifiers is constrained. So, increasing the amount of redundancy means each of them must be smaller. For

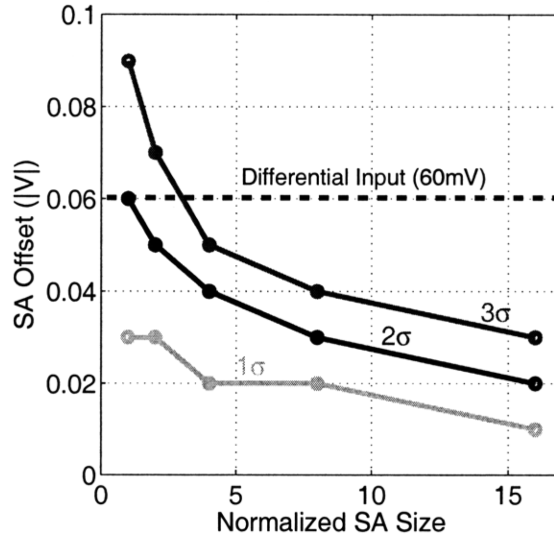


Figure 3-22: Monte Carlo simulations of sense-amp statistical offset; at expected input swing (i.e. 60mV), errors from offset are prominent.

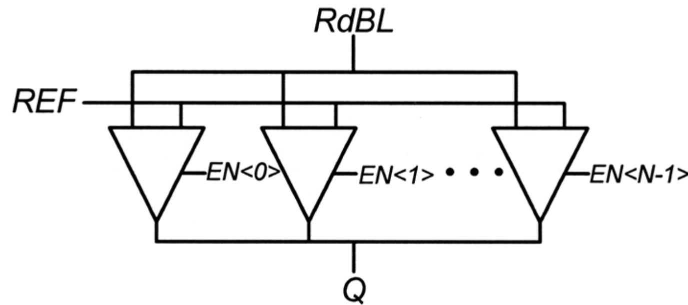


Figure 3-23: With sense-amplifier redundancy, each $RdBL$ is connected to N different sense-amplifiers.

example, as shown in Figure 3-24a, if N equals 2, each must fit into half the allocated area, and, if N equals 4, each must fit into a quarter of the allocated area. Unfortunately, reducing the size of the individual sense-amplifiers in this manner increases the standard-deviation of their offset distribution, and correspondingly increases their probability of error. Specifically, the offset distributions in Figure 3-24b are derived from Monte Carlo simulations of sense-amplifiers designed to occupy a layout active area of $40\mu m^2$, and the error probability for an individual sense-amplifier, $P_{ERR,N}$, is defined as the area under its distribution where the magnitude of the offset exceeds the input voltage swing expected on $RdBL$. Here, it is clear that, due to the necessary

reduction in its size, the error probability for an individual sense-amp increases as the amount of redundancy, N, increases. However, the ability to select one structure with sufficiently small offset means that the error probability for the entire sensing network is the joint probability that all of the individual sense-amplifiers yield an error. The total error probability, $P_{ERR,tot}$, is given by the following:

$$P_{ERR,tot} = (P_{ERR,N})^N \tag{3.1}$$

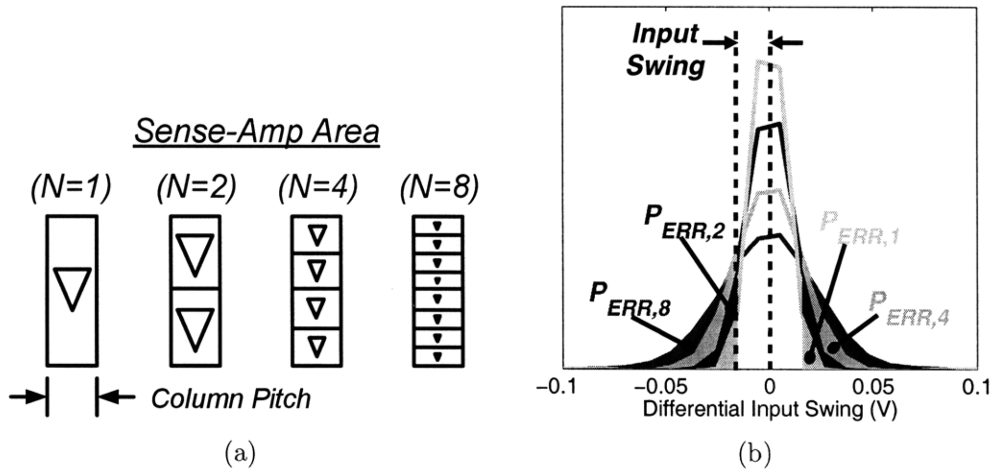


Figure 3-24: With sense-amplifier redundancy (a) the size of each individual sense-amplifier must decrease, and (b) the individual sense-amplifier error probabilities, defined as the area under the offset distribution exceeding the magnitude of the input swing, increases.

The resulting error probabilities for the overall sensing networks are plotted in Figure 3-25 normalized to the error probability of a single, full-sized sense-amplifier. As shown, increased levels of redundancy result in significantly reduced overall error probabilities, and at the input swings expected in this design (i.e. >50mV), the resulting improvement can be well over an order of magnitude. Further analysis of sense-amplifier redundancy, considering its precise area overheads and how it scales with increasing device variability (expected in future technologies), is undertaken below.

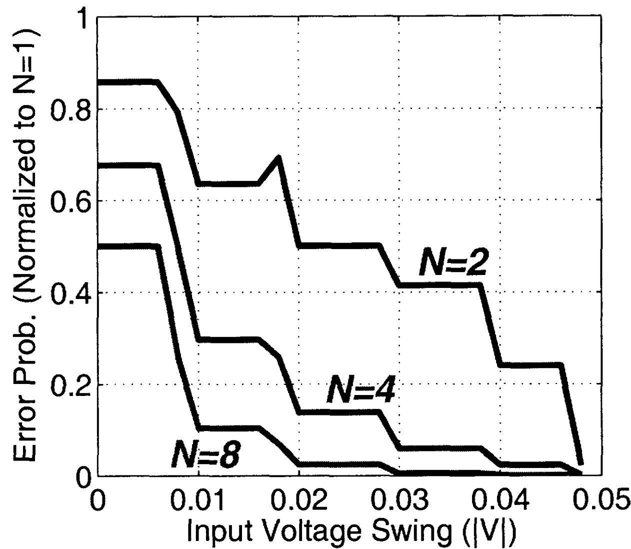


Figure 3-25: Increased levels of redundancy significanly reduce the error probability in the overall sensing network.

Sense-Amplifier Redundancy Implementation

The actual implementation of redundancy used in this design incorporates two sense-amps (i.e. $N = 2$). The analysis above considers a general case of up to $N = 8$, but at those levels, the total area must be large enough to accommodate at least 8 minimum sized structures, and the overhead of the selection logic, which is not considered, becomes significant (the analysis is revised below to consider this overhead). With $N = 2$, the selection logic can be reduced to just two latches and two logic gates.

The rest of the selection circuitry is shown in Figure 3-26. Here, a reference bit-cell is used with both "0" and "1" read-data values hard-wired. This cell gets accessed once on power-up, and it enforces the case where the read-bit-line is first pulled low, and then where it remains high. Fortunately, the logic "1" and "0" levels of the read-bit-lines are fairly independent of variation between the accessed bit-cells; specifically, logic "0" is consistently very near ground, and, as mentioned, logic "1" is set by the aggregate gate, junction, and reverse sub-threshold leakage from all of the read-buffers sharing each read-bit-line. Then, to force the worst-case logic "1" condition, all of the bit-cells in the array must first be written with data that ensures the gate-voltage of

their read-buffer driver-device (i.e. $M7$) is zero, minimizing the reverse sub-threshold current. This can be done in one cycle by simultaneously enabling all word-lines. Consequently, under a static discipline, the wide distribution in read-current does not limit the integrity with which the dummy cell emulates each logic level. Then, the simple state machine in Figure 3-26 determines which of the sense-amps can correctly resolve each logic level, and only the corresponding structure is enabled. If both sense-amps work, the first one is selected, and if neither work, the entire SRAM fails.

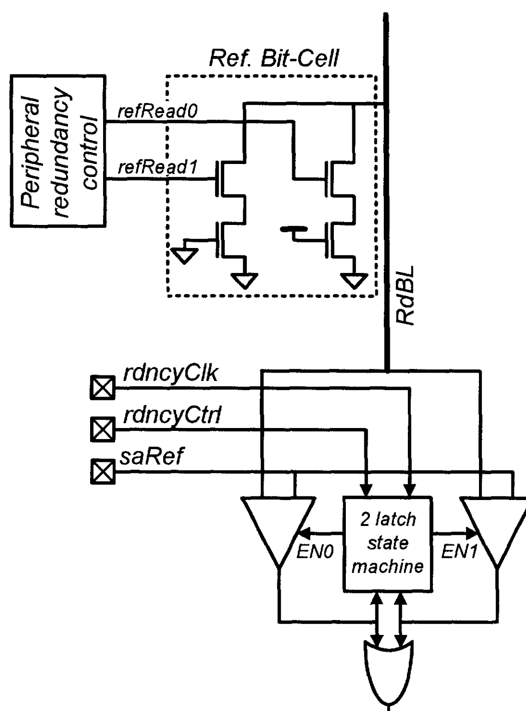


Figure 3-26: Redundancy selection circuitry consisting of a dummy bit-cell and selection state-machine.

Figure 3-27 shows the normalized overall error probability for the sensing-network with the sense-amp sizes actually used in this design. As shown, at the input swings of interest (i.e. ~60mV), the error probability improves by approximately a factor of five compared to a single full-sized sense-amp.

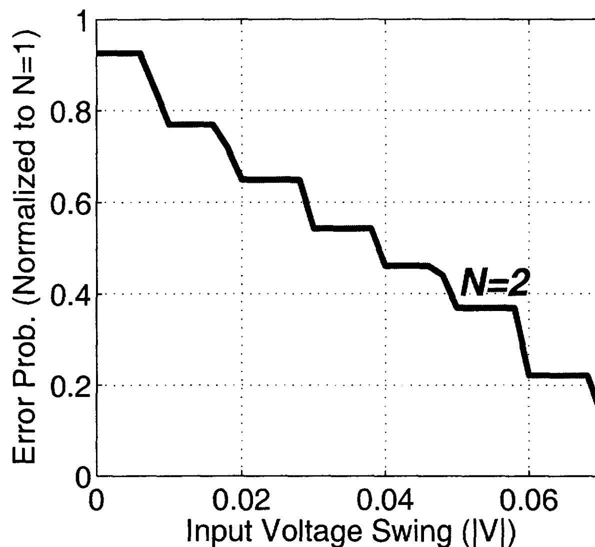


Figure 3-27: Overall error probability for implemented sense-amp redundancy scheme improves by a factor of 5 compared to a single sense-amp scheme.

Redundancy Analysis with Technology Scaling and Overhead

Sense-amplifier redundancy shows promise for easing the area-offset trade-off. However, to evaluate it in a practical scenario, the overhead of the sense-amplifier selection circuitry must be considered. Furthermore, how the benefit of the technique scales with technology is also critical, since it is meant to enable sense-amplifier density scaling at a similar rate as bit-cell density scaling. For this analysis, the case of $N = 2$ (i.e. two sense-amplifiers total) will be considered, since this requires minimal overhead circuitry.

Figure 3-28 shows the overhead selection circuitry required for each sensing-network (i.e. one per $RdBL$). Here, a 6T storage-cell is used to store the state corresponding to the selected sense-amplifier. Then, at power-on, $gSET$ is asserted, initializing the $wrEN$ flip-flop. This allows simultaneous control across all column sensing-networks to selectively enable each of the redundant sense-amplifiers so that all columns can be tested at once using their corresponding hard-wired reference bit-cell. Additionally, however, the $wrEN$ flip-flops from the columns are connected to each other to form a shift-register. Subsequently, this shift-register provides individual and successive assertion of $wrEN$ for each sensing-network so that the desired

sense-amplifier can be permanently enabled on a column-by-column basis. The sense-amplifiers themselves have the structure shown in Figure 3-21, and the input reference, REF, is set to 0.06V, based on the read-bit-line logic levels expected from simulations. Finally, the sense-amplifier ENB signal is driven by a local $NAND$ gate of minimum size. Generally, the ENB transition time can have an impact on the sense-amplifier offset originating from device geometry mismatch and output capacitance mismatch [113]. However, at low-voltages the relative impact of these offset sources is greatly reduced compared to threshold-voltage mismatch, diminishing the need to further reduce the ENB transition time.

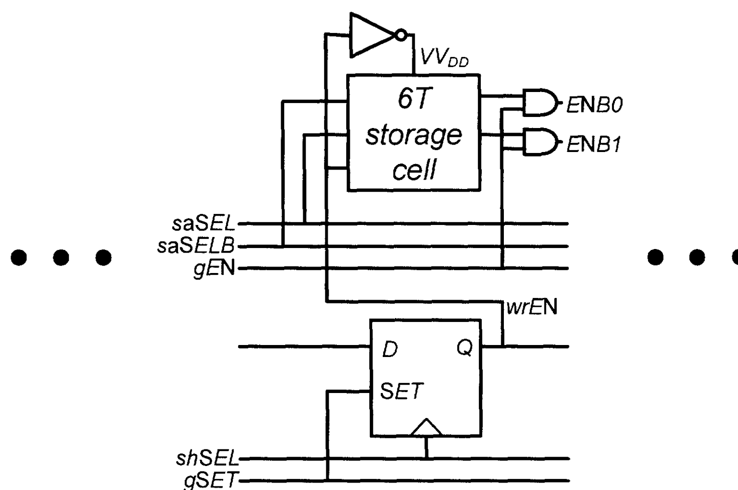


Figure 3-28: Sense-amplifier redundancy overhead circuitry for the case of $N = 2$.

The sensing-networks for both $N = 2$ and $N = 1$, which corresponds to just one full-sized sense-amplifier with no selection circuitry overhead, are layed-out in 65nm CMOS to occupy a total area (A_{tot}) of $40\mu m^2$ each. To consider the impact of technology, scaling, the dimensions of all devices are reduced by $\sqrt{2}$ for 45nm, 2 for 32nm, and $2\sqrt{2}$ for 22nm, and it is assumed that the layout area scales accordingly. For the 32nm and 22nm technology analysis, simulations are performed using a predictive technology model where it is assumed that the threshold-voltage matching co-efficient, A_{VT0}, (i.e. $\sigma V_t \times \sqrt{WL}$) [122] is the same as that of the 45nm devices. Though challenging, in reality, efforts to improve the matching co-efficient are always pursued [123][124][125].

Figure 3-29 shows the overall error probability for the $N = 2$ sensing-network normalized to that of the $N = 1$ sensing-network. The results are derived from 10k point Monte Carlo simulations, and curves are plotted until the input voltage swing exceeds the offset observed from all samples of the respective $N = 1$ sensing-networks (i.e. where their probability of error is too fine for the simulation resolution).

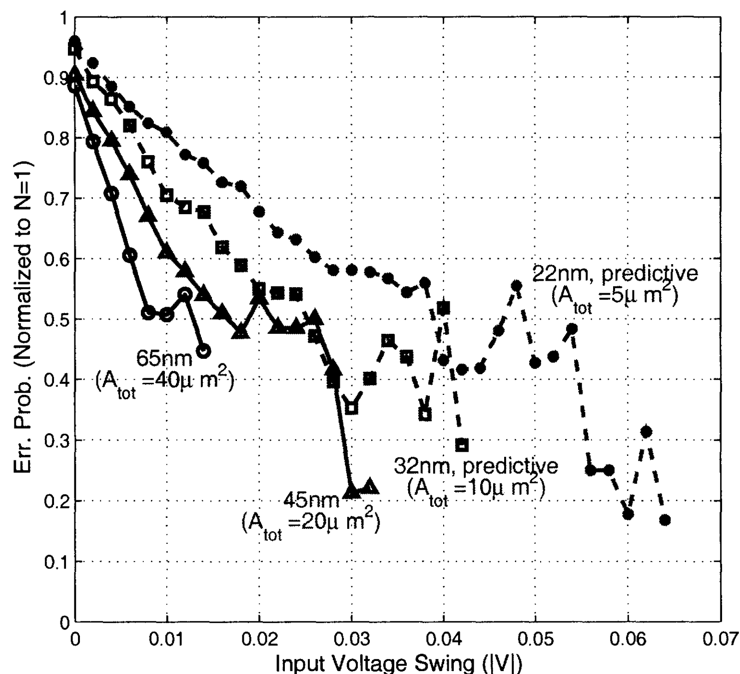


Figure 3-29: Normalized sensing-network ($N = 2$) error probabilities for different technologies and layout areas.

As shown, the selection circuitry overhead, which reduces the active area available for the sense-amplifiers, degrades the benefit of sense-amplifier redundancy somewhat (compared to the result in Figure 3-27). Additionally, the normalized improvement also degrades slightly with respect to technology scaling. Nonetheless, Figure 3-29 shows that sense-amplifier redundancy provides significant improvement in the sensing-error probability, well beyond the 45nm node, even as the total sensing-network area is aggressively scaled from $40\mu m^2$ to $5\mu m^2$.

3.2.3 Test-Chip Architecture

The 256kb SRAM is partitioned into eight bit-cell sub-arrays consisting of 128 columns and 256 rows (32kb), as shown in Figure 3-30. Each cycle, all 128 bit-cells from a row are accessed. No additional delay is required when selecting different sub-arrays, and reads and writes may be performed on consecutive cycles.

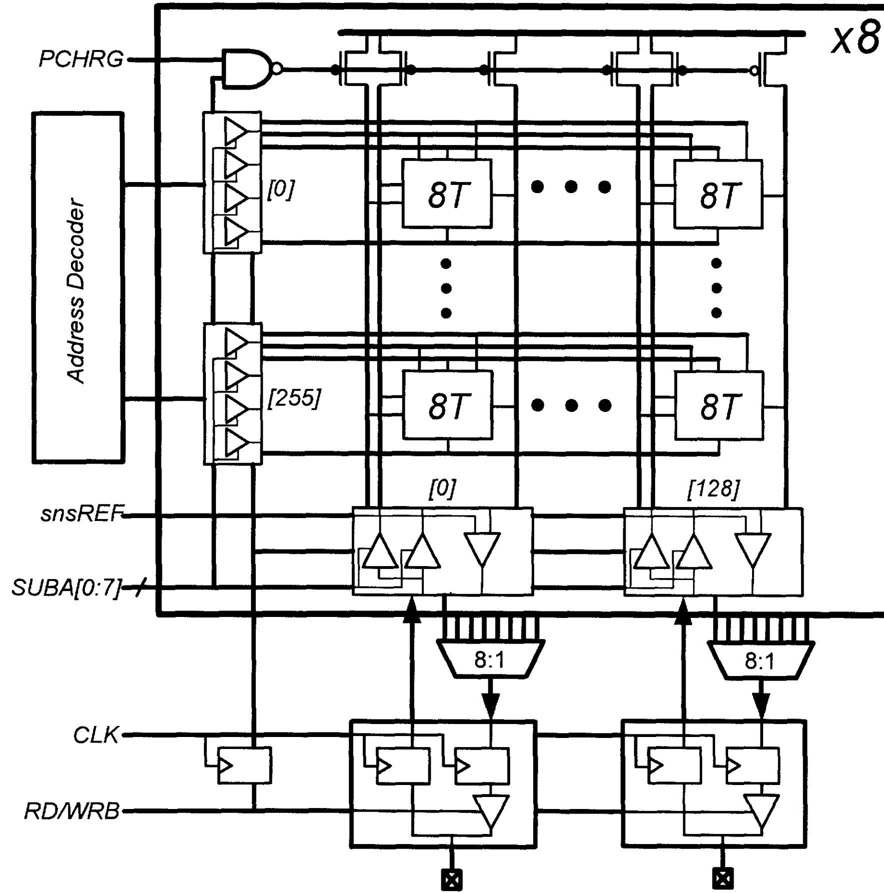


Figure 3-30: Prototype test-chip architecture, with total capacity of 256kb partitioned in eight sub-arrays.

3.2.4 Measurements and Characterization

The prototype incorporating the 8T bit-cell of Section 3.2.1, the peripheral assists, and sense-amplifier redundancy, is implemented in an 65nm LP CMOS technology. A die photograph of the 256kb prototype is shown in Figure 3-31. Measurements are performed by writing and reading two sets of test patterns: (1) checker-board pattern

and its complement over each array, and (2) binary count and its complement down the rows of the array.

The prototype achieves full read and write functionality down to 0.35V (which is well below the device threshold voltages), and it retains data down to 0.3V, indicating that the bit-cell and peripheral assists are successful at enabling a V_{DD} that is close to the data-retention limit. The following sub-sections describe the characterization results of the prototype with regards to its leakage power, active performance, and active power.

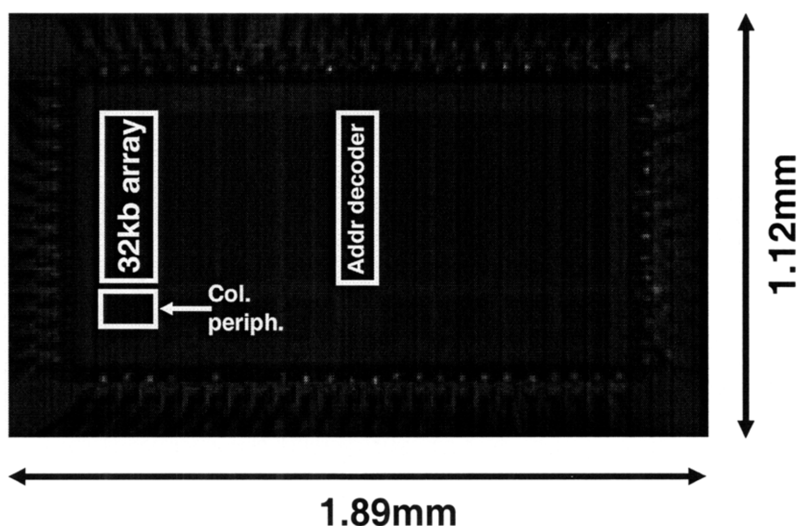


Figure 3-31: Die photo of prototype low-voltage SRAM.

Leakage Power

Figure 3-32 shows the leakage power of the SRAM with respect to supply voltage for $0°C$, $27°C$, and $75°C$. At the minimum V_{DD} of 0.35V, the total leakage power is 2.2μW, representing over a factor of 20 in leakage power savings compared to a supply voltage of 1V. As mentioned, the SRAM also retains data down to 0.3V where the total leakage power is 1.65μW.

The area and leakage power of this SRAM can be compared to a conventional 6T design, and the 10T sub-threshold design in [101]. From the actual cell layouts, this design represents an area overhead of approximately 30% compared to a 6T design

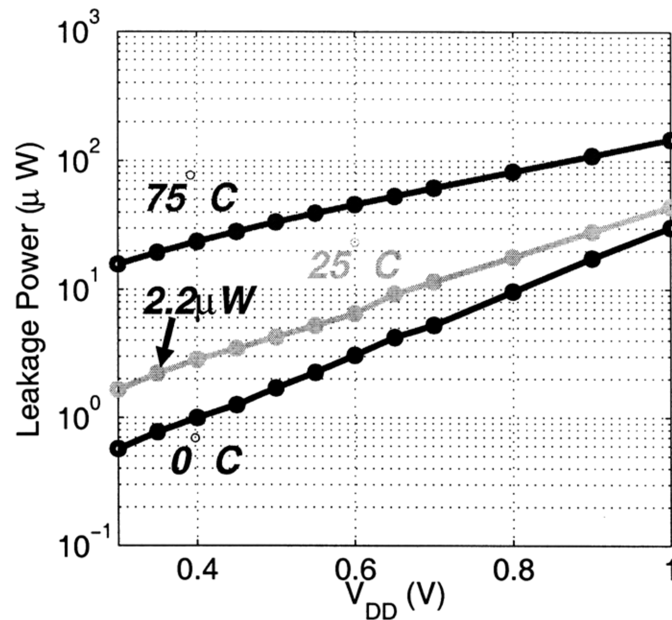


Figure 3-32: Prototype SRAM leakage-power; at the minimum V_{DD} of 0.35V, the entire SRAM draws 2.2μW of leakage-power.

and an area savings of approximately 30% compared to the 10T design. Additionally, the leakage power savings of this design, compared to a conventional 6T design, with a projected V_{DD} of approximately 700mV [57][126], is over 5x.

Active Performance

Figure 3-33 shows the active read and write performance of the prototype SRAM with respect to the supply-voltage. As expected, the speed is significantly reduced in sub-threshold, and at 0.35V, the SRAM operates at 25kHz.

Active Power

Figure 3-34 shows the total (i.e. active plus leakage) power, in the solid curves, and just the leakage power, in the dotted curves, with respect to the operating frequency. The leakage power remains a dominant portion of the total power for a wide range of frequencies, so leakage minimization efforts are well justified.

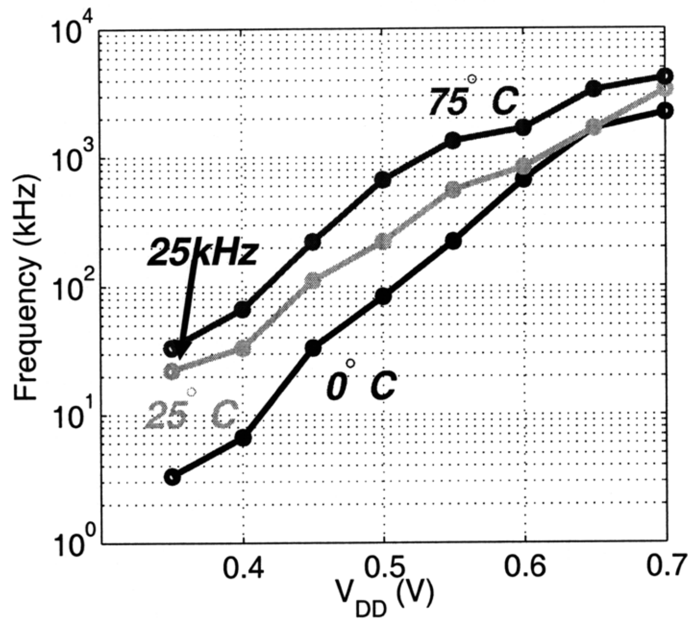


Figure 3-33: SRAM speed with respect to V_{DD}.

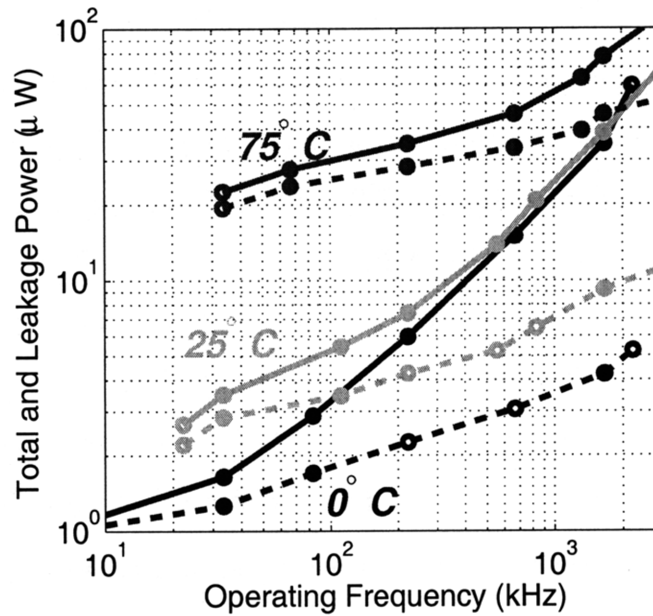


Figure 3-34: Total power (solid curves) and leakage power (dotted curves) with respect to operating frequency.

3.3 Summary and Conclusions

The effects of voltage-scaling on SRAMs are severe. In particular, standard techniques, especially with respect to the bit-cells (i.e. 6T), fall short for reliable operation

at the low voltage (i.e. $< 0.5V$) necessary for highly energy-constrained applications. The severe effect of variation intolerably degrades the read and write margins, though the hold margin is maintained down to the target supply-voltage. Additionally, however, in the presence of bit-line leakage, read-current degradation affects not only performance, but also the ability to sense read-data on the bit-lines.

SRAMs respond much more poorly to voltage scaling than generic logic circuits because the solutions specifically prescribed to enable voltage scaling in logic directly oppose the density and area-efficiency enhancing techniques employed by SRAMs. Accordingly, the simultaneous need to improve density and voltage scalability demands specialized approaches for the bit-cells and sense-amplifiers; the address decoders and word-line drivers, whose density is less critical and whose operation is similar to full-swing logic, can leverage the approaches that are effective for generic logic.

With regards to the 6T bit-cell, the trade-off inherent between read-margin and read-current implies that the viable means to improve the read-margin is through device up-sizing in order to manage variation. However, area comparison in the presence of variation suggests that the 8T topology holds much greater promise for low-voltage operation. Additionally, it increases the options for write-margin, read-current, and sensing-margin (i.e. bit-line leakage) optimization by way of sizing, device-engineering, and selective biasing using peripheral circuit assists. Accordingly, the described prototype employs an improved 8T bit-cell that relies on low-area-overhead peripheral assists to achieve complete operation down to 0.35V.

Finally, redundancy, which is heavily relied upon to manage statistical failures at the 4 or 5σ level in typical bit-cell arrays, is valuable in the periphery when statistical failures, at even the 3σ level, become too significant, as they do at low-voltages. Of course, the benefits of redundancy must be evaluated against the overhead it introduces. Further, its effectiveness must be considered with regards to aggressive density scaling in advanced technologies where variation has increased severity. Analysis shows that redundancy of the sense-amplifiers, which are emerging as a critical limitation to array density scaling, improves their area-offset trade-off significantly

into the 22nm node, despite the associated area overhead.

Chapter 4

Performance Enhancement for High-Density SRAMs

In order to minimize sub-array energy, the analysis in Chapter 2 points to the need to enhance SRAM performance without raising the supply-voltage or reducing the threshold-voltage. Specifically, a reduced active access-period, T_{ACC}, implicitly lowers the leakage-access-energy or, even more significantly, allows further supply- and threshold-voltage reduction, which optimizes the energy for a given performance constraint. Although a variety of techniques, including parallelism and pipelining [4], exist in order to improve the performance of general logic (and promote voltage scaling), SRAM design is highly constrained by its structure and need to maximize density, hence requiring very different approaches. For instance, in high-density, low-power sub-arrays, bit-line discharge (by the accessed bit-cells) poses the critical limitation to access-time, and unlike with multi-stage or parallel data-paths, it cannot be decomposed physically or logically to reduce the overall delay. Further, as described in detail in this chapter, density and noise margin constraints in 6T bit-cells severely limit the options for addressing this critical delay. 8T bit-cells (or other asymmetric topologies) alleviate these limitations greatly, but they increase the complexity of bit-line sensing. Of course, the sense-amplifier itself plays a key role in determining the overall access-time as well, and for both these reasons, it can greatly impact the performance achievable by the sub-array at a given supply- and threshold-voltage.

Importantly, however, sense-amplifiers face their own density-offset trade-offs, which have increased severity in advanced technologies to the point of posing a primary challenge within high-density sub-arrays. Accordingly, this chapter investigates a sensing approach to address sub-array performance while maintaining bit-cell and sense-amplifier density.

Since energy remains the paramount concern, the analysis is undertaken with consideration to low-power technology optimizations. The target technology used to demonstrate the proposed techniques is 45nm LP CMOS, which highlights the variability associated with density maximization. Additionally, the sensing approach is applied to $0.25\mu m^2$ 6T bit-cells, which are the densest achievable in the target technology and further emphasize noise margin and read-current limitations. Since 8T (and other asymmetric) bit-cells hold increasing promise for low-energy SRAMs, the issues associated with single-ended sensing are specifically considered.

The following sections start by identifying the challenges associated with both bit-cells and sense-amplifiers that limit SRAM performance. Then, the advantages of single-ended sensing, both for alleviating the performance limitations and enabling aggressive voltage scaling, are investigated. Finally, the prototype testchip is presented that demonstrates a single-ended sense-amplifier to address the critical performance limitations.

4.1 High-Density SRAM Performance Challenges

Both the bit-cell array and the bit-line sense-amplifiers are critical factors in determining the overall performance of the SRAM sub-array. This section describes the basic trade-offs and challenges inherent in both of these in high-density designs.

4.1.1 Bit-Cell Read-Current

The most urgent performance challenges in bit-cells arise from extreme variation. In the 6T topology this raises an inherent trade-off between read SNM and read-current, severely limiting the sub-array performance improvement achievable by way of bit-cell

design alone. As mentioned, the 8T bit-cell (and several other asymmetric topologies) overcome this trade-off, leading to significant overall performance improvement (as discussed in Chapter 3).

I_{RD}/C_{BL} Degradation

Low-power SRAMs must incur the reduced read-current that comes with technology optimizations to manage leakage-currents, such as raising V_t. Figure 4-1a shows how the read-current of a 6T bit-cell scales further with cell size. The reduction in mean read-current is a direct consequence of reducing the size of the driver devices. However, the increased variation in the smaller devices also results in more severe degradation to the weak-cell read-current in proportion to this already reduced mean read-current. For instance, at the nominal voltage of 1.1V, the 5σ read-current for the 6T $0.25\mu m^2$ cell is easily degraded by an additional factor of two, and, as mentioned in Chapter 3, the degradation is even more severe as supply-voltage is reduced.

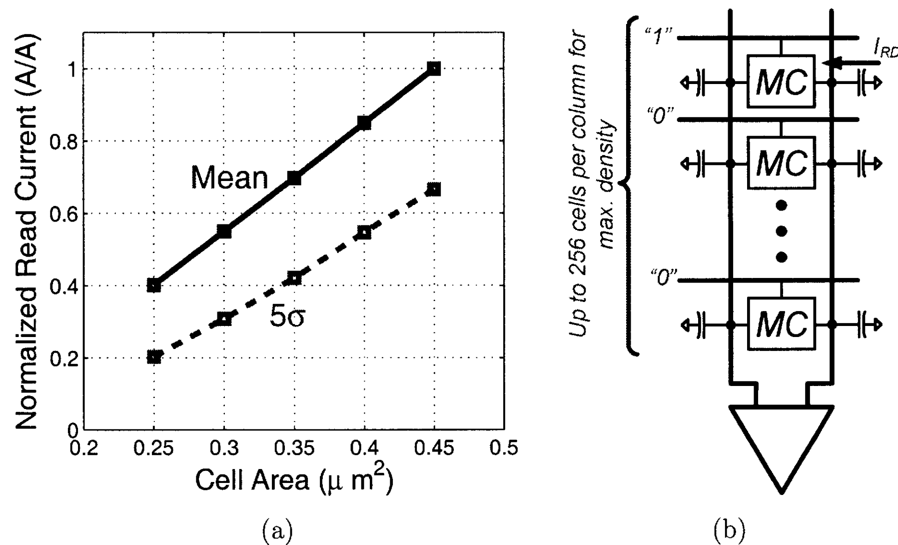


Figure 4-1: Degradation in bit-line discharge time for high-density SRAMs caused by (a) reduced cell read-current and (b) increased bit-line capacitance.

Furthermore, in high-density arrays, the integration of many cells per bit-line (up to 256) in order to maximize array area-efficiency, leads to very large bit-line capacitance. As shown in Figure 4-1b, the resulting ratio of $\frac{I_{RD}}{C_{BL}}$, which is critical to the array's performance, suffers even further.

6T Read SNM Degradation

In 6T cells, increased variation in the highest density bit-cells degrades the read SNM in the same manner as the read-current [79]. Figure 4-2a shows this reduction (at the 5σ level) as the cell area is scaled. Once again, this result is at the nominal voltage of 1.1V, but the impact is even more severe with supply-voltage scaling. Thus, in high-density 6T bit-cells, read-current and read SNM, which are both critical metrics, are simultaneously stressed. Exacerbating the matter even further, these metrics exhibit a strong inverse correlation, as shown in Figure 4-2b [127]. As mentioned briefly in Chapter 3, this comes about as a result of the bit-cell access-devices, which must be strong for maximum read-current but weak for maximum read SNM. Unfortunately, then, optimizations targeting one are likely to worsen the other. Nonetheless, for functionality, read SNM remains the paramount concern, and cell design trends and circuit assists preferentially aim to improve it, leading to further reduction in read-current [39].

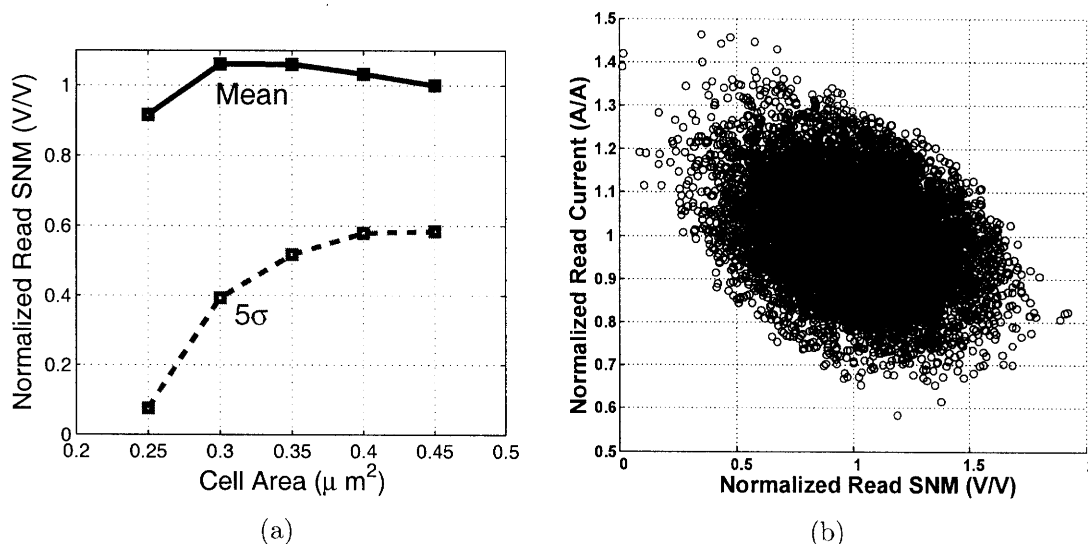


Figure 4-2: Read SNM trade-off in high-density SRAMs limited by (a) cell size and (b) inverse correlation with cell read-current, caused by opposing access-device requirements.

Density and stability trade-offs imply the urgent need for an alternate method to recover the ensuing loss in array performance. As mentioned in Section 4.2, 8T bit-cells (which overcome this interdependence), impose the additional need for single-

ended sensing.

4.1.2 Sense-Amplifier Delay and Uncertainty

The total read access-time depends on both the bit-line discharge delay and the sense-amplifier delay. This section considers the factors affecting the sense-amplifier delay and those related to the sense-amplifier offset; offset sets the required bit-line discharge, thereby strongly affecting the overall delay. Further, just like offset is managed by increasing the timing margin, strobing timing variation introduces timing uncertainty, raising the need for additional timing margin and further limiting the overall delay.

Sense-Amplifier Delay

Two commonly used strobed sense-amplifiers are shown in Figure 4-3. The delay of these structures, after assertion of the $STRB$ signal (whose own impact is considered in detail below) can be separated into two phases [128]: 1) development of output differential, and 2) regeneration of output. Though output differential begins developing (somewhat regeneratively through the NMOS devices) immediately after $STRB$ assertion, full regeneration is defined to begin after the PMOS loads turn-on, actively pulling-up one of the output nodes.

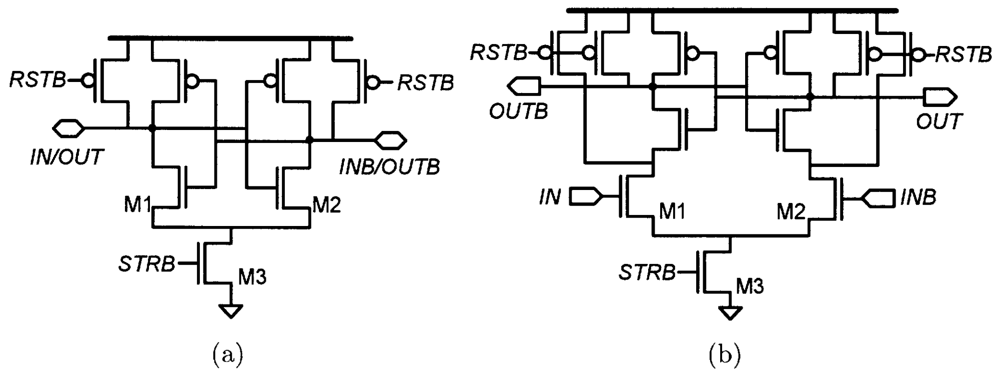


Figure 4-3: Conventional strobed sense-amplifier topologies with (a) one input-output port and (b) separate input-output ports.

Output differential (which is also the input differential in the case of the first

sense-amplifier) is developed through the NMOS pull-down devices ($M1-3$). As a result, it depends on the initial input differential and the NMOS transconductances. Of course, the initial input differential, after full assertion of the word-line, depends on the bit-cell read-current and the bit-line capacitance. On the other hand, the delay to regenerate the outputs to restored logic levels depends primarily on the initial output differential, which is amplified at an exponential rate. Consequently, a large inital output differential reduces the delay of the rail-to-rail output regeneration phase. However, as expected, techniques to generate a large output differential can considerably extend the first phase, and, as a result, the overall optimal conditions are highly dependant on implementation parameters, including the rate of bit-line discharge and the sense-amplifier structure used.

For instance, reducing the input common-mode voltage, $V_{IN,CM}$, reduces the strength of both NMOS pull-down paths, leading to increased time for output differential development; however, the proportional reduction in strength of the pull-down path associated with the drooping bit-line is increased, resulting in larger output differential at the start of the second phase. Hence, rail-to-rail regeneration occurs more quickly. Accordingly, in [128], the optimal delay for the sense-amplifier considered is achieved with an initial bit-line precharge voltage at approximately 80% of V_{DD}. It is important to note, however, that although slightly reducing the bit-line precharge voltage can improve the read SNM, it degrades the bit-cell read-current (with increasing impact in advanced technologies).

Reducing the transition rate of the $STRB$ signal has a similar effect to reducing the bit-line precharge voltage. It, once again, reduces the strength of both NMOS pull-down paths, increasing the delay of the first phase (i.e. output differential development) but also resulting in a larger output differential at the end of this first phase. In [113], the optimal delay for the entire access-time is achieved by slowing the $STRB$ edge delay to over twice its minimum transition time.

Sense-Amplifier Offset

Sense-amplifier offset leads to an increase in the bit-line droop required in order to ensure correct data read-ability, extending the read access-time. The trade-off between device up-sizing and offset is well known, both with regards to area, which affects V_t variation [119][120], and linear length and width, which affects geometry variation [122]. This trade-off is the primarily limitation to sense-amplifier area scaling, which is emerging as a dominating limitation in high-density sub-arrays [111].

The bias conditions of the sense-amplifier devices also strongly impact its overall offset in the presence of variation. For instance, during the first phase (i.e. output differential development), lowering the gate overdrive ($V_{GS} - V_t$) of the input devices ($M1 - 2$) reduces the relative impact of geometric and load-capacitance imbalance between the two sense-amplifier branches compared to the intended input voltage imbalance. As a result, offsets due to geometry and capacitance variation lead to reduced input offset [129][130]. This serves as motivation to minimize the edge rate of the $STRB$ signal, which is shown to reduce offset by over 40% [113]; and it serves as motivation to reduce the bit-line precharge voltage, which improves the sensing yield by over 30% (for a 60% reduction in the precharge voltage) [128].

In addition to sense-amplifier offset mitigation through device sizing and biasing, various forms of active offset cancellation have also been investigated. For instance, the DRAM designs in [131] and [131] actively bias the bit-lines to compensate the sense-amplifier's input offset. Since the bit-line capacitance is quite large, these schemes can lead to long offset-compensation phases and high power consumption. Alternatively, the offset can be compensated by triming the discharge rate of of the sense-amplifier's differential branches; using switchable current-sources and load-capacitors, this can be accomplished with a highly digital scheme [132][133], though an explicit calibration phase and control circuitry is necessary.

Sense-Amplifier Strobing Uncertainty

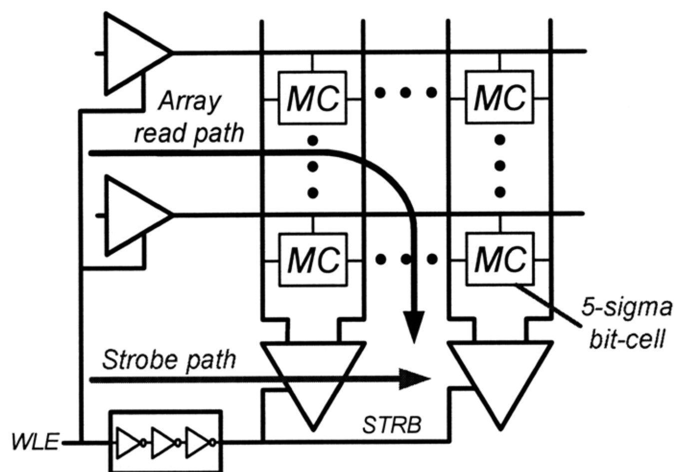
The sense-amplifier strobe signal must ensure that enough time is allocated for bit-line discharge to overcome the sense-amplifier offset. However, as shown in Figure 4-4a, the bit-line discharge time depends on the delay through the array read-path, which is limited by the weakest bit-cell. Such a cell, whose statistical characteristics might be beyond the 5σ level, is impossible to replicate in the strobe control path. Consequently, as shown in Figure 4-4b, the timing of the two paths diverges greatly over process, voltage, and temperature corners [134], even if carefully laid-out SRAM devices are employed in the strobe path, as in [135].

For this simulation, the read-path is designed to overcome a sense-amplifier offset of 50mV, and an array configuration of 256×256 is considered; however, the strobe path must be designed such that it is longer than the array read-path in all cases. As shown in Figure 4-4b, this implies that in many cases it will be much longer than it needs to be, thereby excessively limiting the overall performance. In fact, the overall worst-case delay need only be 820ps based on the read-path delay; however, the actual worst-case delay is 980ps, limited by the strobe-path, imposing an excess overhead of nearly 20%. To recover this, designs such as [134] have gone as far as using temperature sensors to adjust the relative read-path and strobe-path delays, and automatic compensation based on process and voltage can be even more challenging.

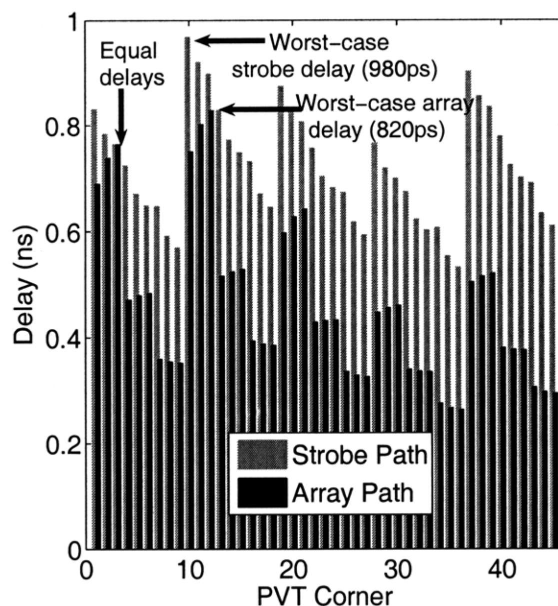
4.2 Single-Ended Sensing

With regards to the bit-cell, one way to address the severe read SNM limitation and independently improve the read-current is to use an alternate structure. Asymmetric cells, including 6T [97], 7T [127], and 8T [93] topologies, can have much wider operating margins and significantly increased read-current, making them particularly compelling in the face of increasing variation [103][13][136]. However, none of these provides a differential read-port, and, therefore, they require a compatible means for efficient single-ended sensing.

The most common technique used with single-ended-read bit-cells is full-swing



(a)



(b)

Figure 4-4: Array read-path and sense-amplifier strobe-path (a) limited by matching to 5σ bit-cell and (b) exhibiting severe delay divergence over process-voltage-temperature conditions, leading to excess overall delay.

sensing [111]. Here, the read bit-line must discharge almost completely so that the read data can be detected by a sensing structure that may be reduced to essentially a logic gate. This, however, leads to excess active-switching energy, and, in order to maintain performance, bit-line discharge times must be reduced by enhancing the read-current and minimizing the bit-line capacitance. Since altering supply- and threshold-voltages impacts the energy, this implies the need for large bit-cells and very

short bit-lines. Unfortunately, both of these directions severely degrade the density.

Alternatively, small-signal sensing can be retained by employing a pseudo-differential sense-amplifier, as in [137]. However, the need to generate a complementary reference is unavoidable, and, importantly, it must track all operating conditions even with respect to cells in the tails of the arrays statistical distribution. As a result, design and testability of the complementary reference is highly complicated.

Noise Sensitivity

An important draw-back to any single-ended sensing scheme is the loss of common-mode noise rejection capability on the power-supplies, bit-lines, and substrate. Since no distinction can be made between noise on the bit-line and read-data droop, it becomes critical to ensure a desired level of bit-line noise-margin. Consequently, an inherent trade-off is introduced between sense-amplifier input sensitivity, which ultimately affects the sub-array's performance, and its robustness to noise. For power-supply and substrate noise, though there is no inherent trade-off between sensitivity and noise-rejection, it is critical to characterize the sense-amplifier's noise margins. This can be extremely challenging since, in general, phase and frequency content of the transient noise can change its impact on the sense-amplifier as they do for bit-cells [138].

4.3 High-Density SRAM Prototype

In this section, a sense-amplifier designed to maximize the performance of high-density low-power arrays is presented. It is prototyped in an 45nm LP CMOS process and integrated with a 64kb array composed of $0.25\mu m^2$ bit-cells. The high-density bit-cells and low-power technology optimizations stress read-current limitations, and the array configuration (256×256) stresses read-bit-line capacitance, which all lead to long discharge times.

4.3.1 Non-Strobed Regenerative Sense-Amplifier

The non-strobed regenerative sense amplifier (NSR-SA) shown in Figure 4-5 addresses the limitations described in Section 4.1. Two cascaded inverters ($M1-2$ and $M3-M4$) form an amplification path that is self-biased for high-gain by the feedback switches, S_{AZ}. As described below, these also perform offset-compensation of the inverter amplifiers via input auto-zeroing [139]. Importantly, to support small-signal sensing, which greatly improves the density, performance, and power trade-offs of the array, the large gain required is achieved most efficiently through regeneration [140]. Accordingly, the regenerative device, $M5$, provides very large positive-feedback gain. A critical feature is that regeneration requires no external enable or strobe signal, thereby overcoming the severe tracking uncertainty described in Section 4.1.2. Instead, the ideal DC transfer function, shown in Figure 4-5, is implemented in continuous time, and the precise point at which regeneration is enabled depends only on the input bit-line voltage (with respect to the internal reference, V_{TRIP}). Further, V_{TRIP} is very stable despite variation since it is generated implicitly by the original auto-zeroing, and, additionally, its precise value can be selected by design. Accordingly, this sense-amplifier, which is single-ended and compatible with asymmetric bit-cells, can enforce a desired balance between sensitivity and noise-rejection.

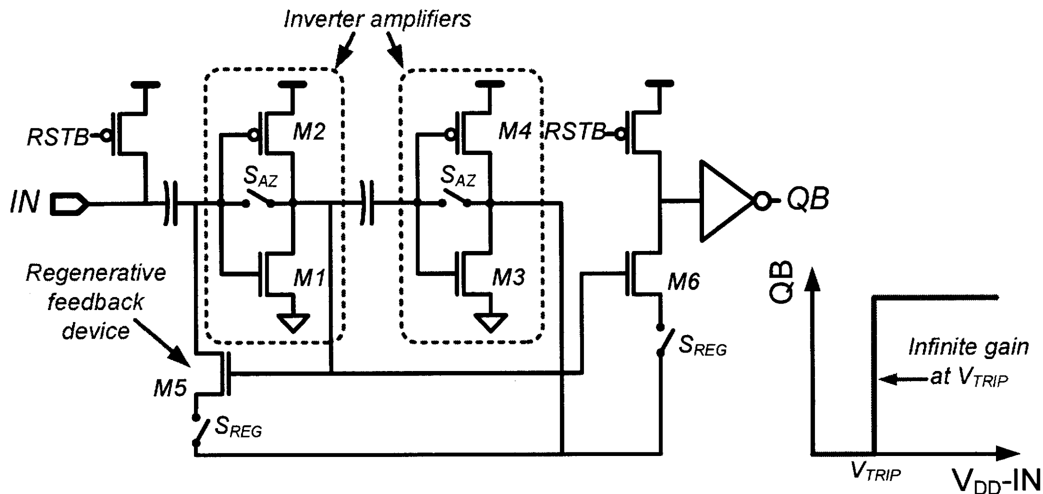


Figure 4-5: Non-strobed regenerative sense-amplifier (NSR-SA) schematic and ideal transfer function.

Basic Operation

The NSR-SA operates over two phases: reset and detection. The reset phase occurs during SRAM bit-line pre-charge, where the sense-amplifier does not need to detect bit-line droop. Hence, this time is used to perform self-correction of offsets via auto-zeroing. The detection phase corresponds with bit-line discharge, where the actual read-data must be quickly resolved. Details of the reset and detection phases, as well as how they overcome the need for output-register clocking-margin, are provided below:

(1) **Reset Phase.** The purpose of the reset phase is to charge the internal nodes so that the inverters formed by $M1-2$ and $M3-4$ are biased in their high-gain regions very close to their ideal trip-points. Simultaneously, this initializes the regenerative device, $M5$, such that its positive feedback gain is very low.

Although the reset phase occurs during bit-line pre-charge, it is actually meant to occupy only a small portion of this period. For instance, as shown in Figure 4-6, a short RST pulse is asserted (even its duration as shown is much longer than required). During this time, the input node and output stage ($M6$) are pre-charged, while, simultaneously, the negative feedback switches, S_{AZ}, are closed and the regeneration switches, S_{REG}, are opened. Consequently, as shown in Figure 4-6, nodes X and Y get biased to mid-rail voltages at the inverter trip-points, which are nominally designed to be equal. Offsets can lead to differences in their precise value, but the offset compensation analysis describes how this biasing greatly attenuates the errors.

It can be seen that nodes X and Y settle to their required reset values in less than 100ps, implying that only a very small RST pulse is required. It should be noted, however, that while nodes X and Y remain at these mid-rail voltages, a static current path exists through $M1-2$ and $M3-4$. As discussed in Section 4.3.3, however, the total power overhead introduced as a result is small.

(2) **Detection Phase.** Following reset, bit-line discharge must be detected. First, the case where the bit-line remains high at its pre-charge value (i.e. logic

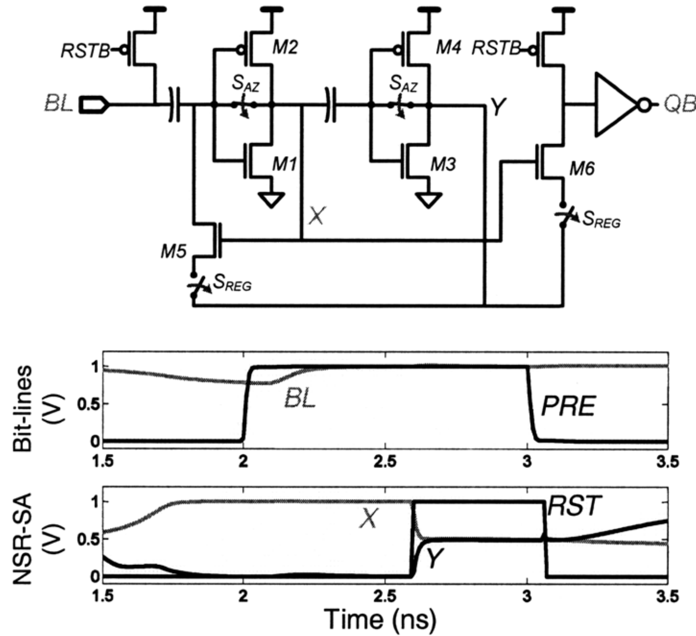


Figure 4-6: NSR-SA circuit and waveforms during reset phase.

"1") will be considered. Here, the bit-line voltage, BL, remains unchanged, so all internal voltages remain essentially at their reset bias values. For instance, as shown in Figure 4-7, after the negative feedback switches, S_{AZ}, are opened, nodes X and Y remain unchanged aside from small perturbations arising from charge-injection errors from the S_{AZ} switches (which will be addressed below with consideration to false regeneration immunity). Accordingly, when the S_{REG} switches are closed to enable $M5$ and $M6$, the V_{GS} of these devices remains very small; in fact, it is even designed to become slightly negative thanks to charge-injection errors originating from the specific choice of the devices used to implement the S_{AZ} switches (this is discussed further later). As a result, the output logic level of $QB = 0$ is sustained. Of course, bit-line leakage arising from the unaccessed cells sharing BL can compromise the logic "1" value, leading to detection errors. Although the NSR-SA's BL noise-margin can be designed to reject these (as discussed with regards to the regeneration trip-point design), maximum-leakage simulations in the target LP process indicate that, for the 256 cells/BL configuration considered, the nominal NSR-SA rejects bit-line leakage for access-times up to 35ns, which are much longer than the target delays.

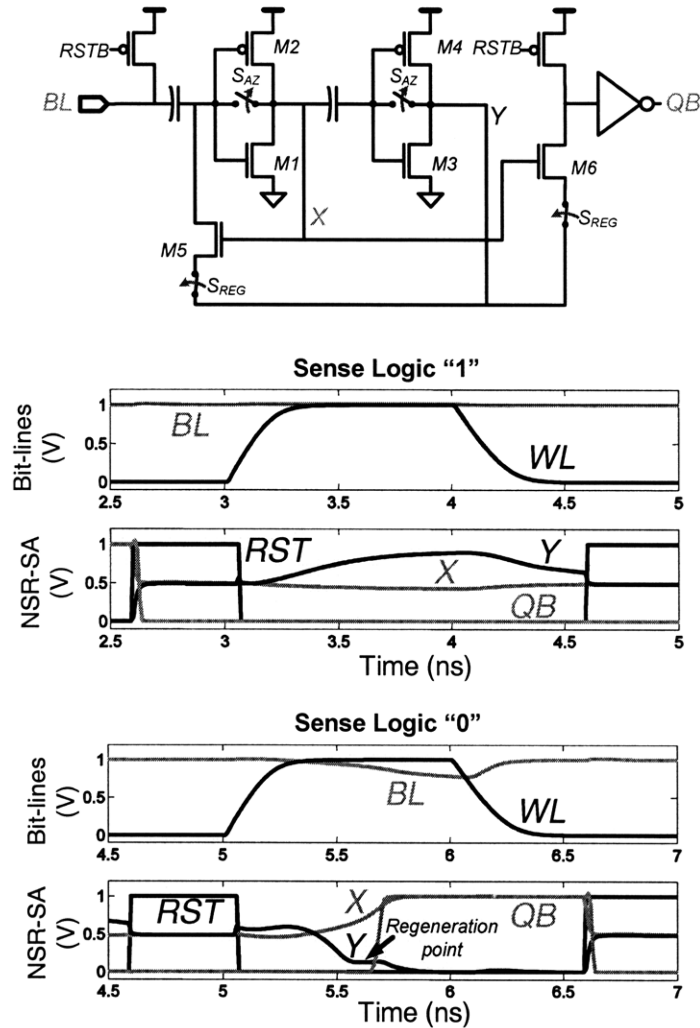


Figure 4-7: NSR-SA circuit and waveforms during detection phase (for both bit-line logic cases).

Alternatively, in the logic "0" case, also shown in Figure 4-7, an intentional bit-line droop is detected. Here, the voltage of node X rises rapidly as a result of the inverter gain, and the voltage of node Y decreases even more rapidly as a result of the cascaded inverter gains. Correspondingly, at some point the regenerative device's (i.e. $M5$'s) V_{GS}, which is the difference between the voltage of nodes X and Y, becomes large enough that the device turns on, triggering positive feedback. Subsequently, the input of the first inverter is actively pulled low, and the entire NSR-SA quickly latches. The precise point where regeneration is enabled can be seen at the annotated inflection in the waveform of node Y.

Shortly after this, the output, QB, quickly changes its state.

(3) **Output Clocking Margin.** As mentioned, an important feature of the NSR-SA is that regeneration is triggered by the input bit-line droop itself, rather than an explicit strobe signal. Nonetheless, the read data must ultimately be clocked at the array output. However, as shown in Figure 4-8a, the timing problem described in Section 4.1.2 has actually been overcome and not just propagated to the subsequent output clock, $OCLK$.

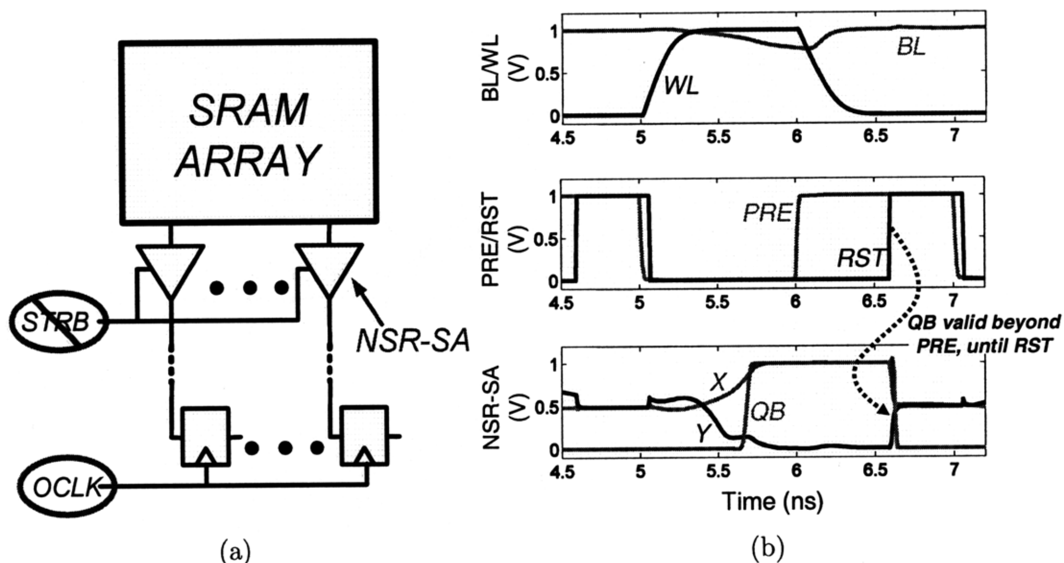


Figure 4-8: Output clocking (a) at array-level with (b) waveforms showing decoupling from internal critical read-path.

By comparison, for instance, the strobe signal of a conventional sense-amplifier must arrive during the bit-line discharge phase, before the next pre-charge phase can begin. As a result, the timing of the strobe signal limits the critical read-path inside the array. However, as shown in Figure 4-8b, the NSR-SA latches the output without a strobe signal, and the output state remains valid until the reset pulse is asserted. Since the reset pulse can be a small fraction of the pre-charge phase, the output data can be clocked well after the pre-charge phase begins, until the data is cleared by the reset pulse. As a result, the timing of $OCLK$ is decoupled from the critical read-path inside the array.

Offset Compensation

For reliable small-signal sensing, stability of the input trip-point, V_{TRIP}, is critical. This is achieved by the autozeroed biasing enforced by the S_{AZ} switches. To analyze the effect of these switches, Figure 4-9a abstracts the inverters of the NSR-SA by their logic symbol. Each of these inverters is modeled as being ideal and offset-free, and the offsets are modeled as the series voltage sources, V_{OS1-3}, which are associated with each inverter and the regenerative device.

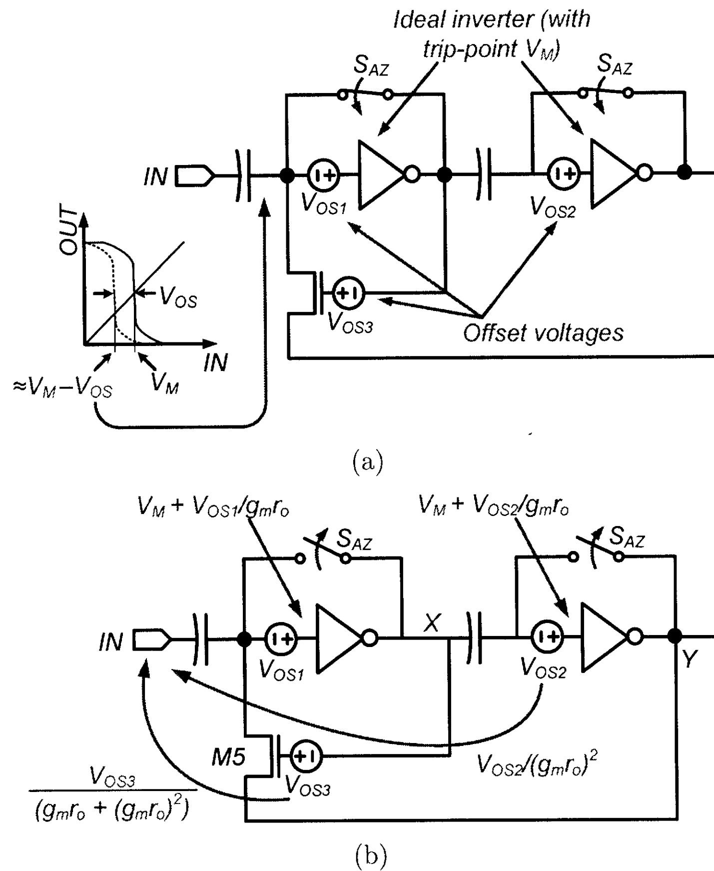


Figure 4-9: Offset compensation (a) technique and (b) analysis.

During the reset phase, the S_{AZ} switches are closed, forcing the inverters into negative feedback. Analytically, this implies that the input and output values of the voltage transfer characteristic (VTC) must be equal, as shown by the diagonal line. Accordingly, if the inverter is offset-free, as in the case of the solid VTC, its input would settle to the voltage V_M, which is the ideal trip-point. However, offset shifts the VTC to the left by an amount equal to V_{OS}. Now, as shown by the dotted VTC, the

input instead settles to a value approximately equal to $V_M - V_{OS}$. In particular, this requires that the VTC is nearly vertical near these trip-points, which corresponds to a large inverter gain. If this is the case, the resulting voltage of $V_M - V_{OS}$ gets stored on the preceding coupling capacitor. As a result, $-V_{OS}$ appears in series with the actual V_{OS}, canceling the offset-voltage and biasing each inverter to its ideal trip-point.

Complete offset cancellation in this manner, however, requires that the gains of the inverters be infinite. Practically, the offset is only reduced by a factor equal to the finite inverter gain [139], which is approximately given by $g_m r_o$, where g_m is the transconductance and r_o is the output resistance of the inverter devices. As shown in Figure 4-9b, however, the residual offset of the second inverter stage is reduced by an additional factor of $g_m r_o$ when input referred, which requires dividing by the preceding inverter's gain. Finally, nothing explicit is done to manage the offset of the regenerative device, but, once again, its contribution to the overall offset is very small when input referred, since the gain from the input to its V_{GS} is given by $g_m r_o + (g_m r_o)^2$ through the inverter path. Accordingly, after all of the attenuation factors in Figure 4-9b, the offset of the first inverter dominates, but it is significantly reduced thanks to the offset-compensation.

A Monte Carlo simulation of the access-time distributions for the NSR-SA and a conventional strobed sense-amplifier is shown in Figure 4-10. In this context, access-time includes the delay of the word-line driver, bit-line discharge, and sense-amplifier, and it is measured from the word-line enable signal to the sense-amplifier output valid. A 256-by-256 array configuration is considered with a mean bit-cell, so that the variation in the sense-amplifiers can be isolated. The conventional sense-amplifier has been sized for minimum offset while occupying a layout area of $12\mu m^2$. In order to determine the appropriate delay of its strobe signal, two factors are considered: (1) the bit-line discharge required to overcome its offset and generate the correct output within 200ps of strobe assertion, and (2) the timing-divergence of the array read-path (as described in Section 4.1.2). Simulations of the standard strobed sense-amplifier show that with an input much larger than the offset, approximately 70ps is required after the arrival of the strobe signal in order for the output to regenerate to the correct

state. Hence, the strobe signal is designed to allow enough bit-line discharge in order to ensure a sense-amplifier delay less than 100ps. An additional delay of 160ps is also inserted, based on process, voltage, and temperature (PVT) corner simulations, to account for read- and strobe-path timing divergence over operating conditions. Finally, the transition time of the strobe signal is set to 100ps.

As shown, the standard strobed sense-amplifier achieves good mean performance thanks to its differential operation, as, nominally, it must wait long enough for only a very minute differential to develop on the bit-lines. However, because it is more sensitive to variation, its sigma is worse at 37ps, compared to 18ps for the NSR-SA, and, therefore, it achieves slower worst-case performance.

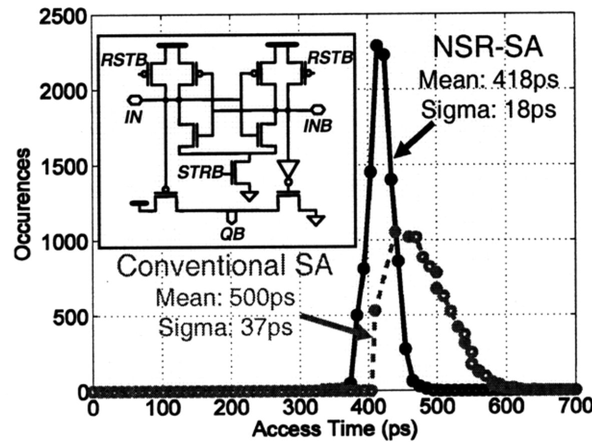


Figure 4-10: 10k point Monte Carlo simulation showing improved sigma of NSR-SA access-time compared to conventional sense-amplifier access-time.

In addition the statistical access-time distribution shown, the NSR-SA offset compensation can also be analyzed by means of an input-output voltage transfer characteristic in order to evaluate the stability of its trip-point, V_{TRIP}. This analysis is provided below, with consideration to noise-margins. Finally, the improved stability achieved by the NSR-SA points to the benefit of offset compensation, but the next limitation to its sigma comes from variation in the charge-injection errors introduced by the S_{AZ} switches.

False Regeneration Immunity

Apart from residual standard deviation in the delay, an even more urgent failure mode arising from charge-injection errors is that, in an amplifier with high gain and regeneration, these can potentially result in resolution to the wrong output state, from which recovery might not even be possible. For instance, in the NSR-SA, if charge-injection errors were to cause the input of the first inverter to decrease and the input of the second inverter to increase, the gain through the inverters would cause a large positive V_{GS} on the regenerative device, $M5$, causing the output to transition and latch, regardless of the input bit-line voltage.

The NSR-SA exploits the fact that it only needs to respond to bit-line discharge, not up-charge. As a result, regeneration only needs to occur in one direction, and the charge-injection error sources can be designed to oppose that direction, as shown in Figure 4-11. Specifically, the main device used for the first S_{AZ} switch is a PMOS, whose charge-injection errors tend to increase the input of the first inverter, and the main device used for the second S_{AZ} switch is an NMOS, whose charge-injection errors tend to decrease the input of the second inverter. As a result, nodes X and Y decrease and increase respectively, giving a negative V_{GS} on $M5$, pushing it away from false regeneration. While false regeneration can be avoided in this manner by proper design of the S_{AZ} switches, the effect of charge injection-error variation on the NSR-SA's trip-point cannot. This is considered in more detail below with regards to the noise margins.

Regeneration Trip-point

Although offset-compensation improves the stability of the NSR-SA in the presence of variation, it is also important to set its nominal trip-point, V_{TRIP}, based on speed and noise-rejection considerations. One way to achieve this control is by adjusting the reset voltages of nodes X and Y, which change the V_{GS} of $M5$ and trim the amount of additional bit-line discharge required to actually trigger regeneration. This can be implemented, for instance, with the addition of an appropriately sized device at

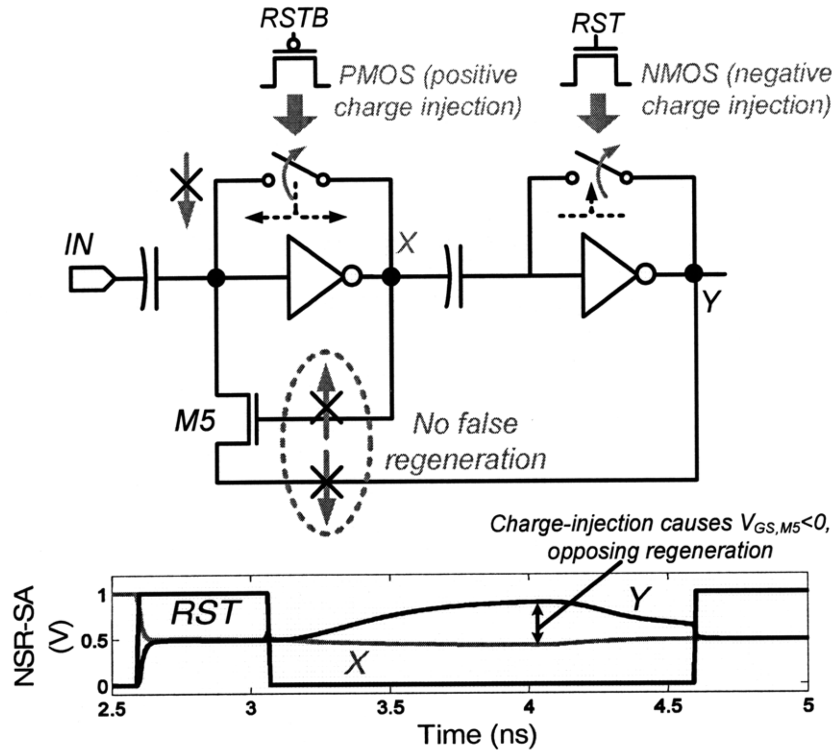


Figure 4-11: NSR-SA robustness to false-regeneration in the presence of charge-injection errors.

the inverter outputs, as shown in Figure 4-12. This device can be either an NMOS or PMOS (NMOS is shown) depending on the trimming required, and, with its gate connected to an appropriate rail voltage, even a very small device (less than $0.2\mu m \times 0.2\mu m$) results in a wide range of output reset voltages (covering almost 0.3V in this case).

Noise Margins

As mentioned in Section 4.2, noise on V_{DD}, V_{SS}, BL, and substrate all compromise correct data-sensing by the NSR-SA. In particular, noise at these nodes can result in false regenerations, which will be called output errors, and deviation from the intended nominal input trip-point (V_{TRIP}), which will be called input errors. As before, false regenerations are the primary concern, since they directly imply erroneous sensing and cannot be recovered from. However, in the presence of BL noise, input errors also lead to erroneous sensing and degrade sensitivity by increasing the amount of

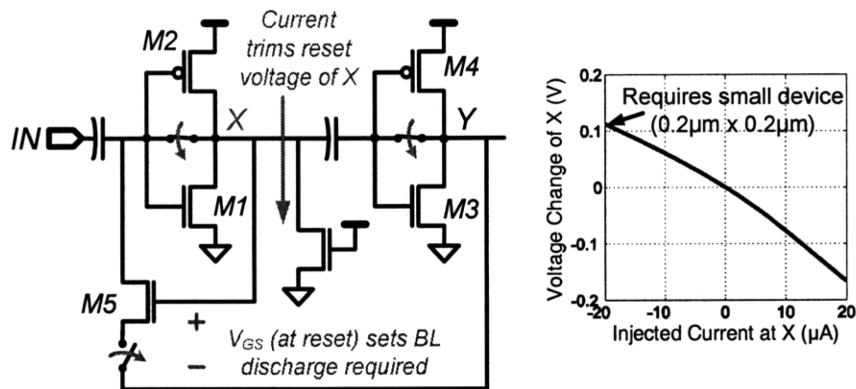
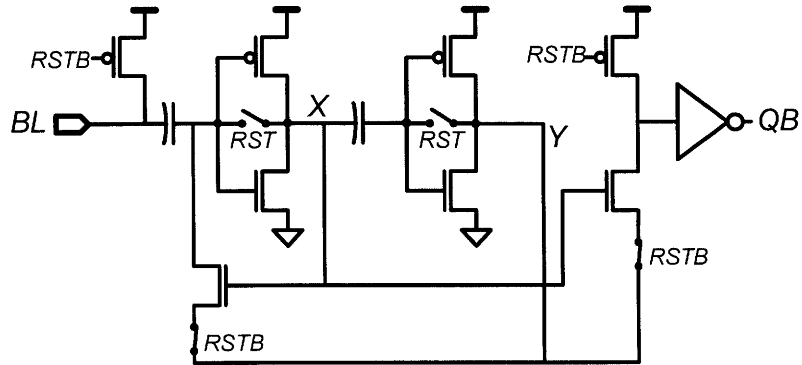


Figure 4-12: NSR-SA technique to set regeneration trip-point (V_{TRIP}) for noise-rejection and sensitivity considerations.

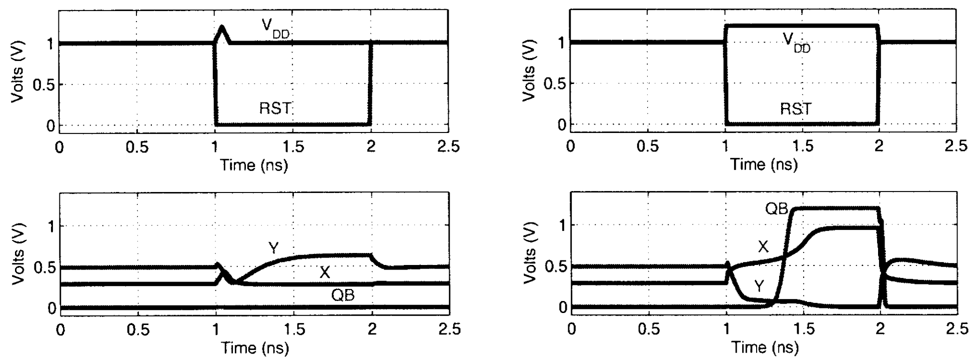
nominal bit-line noise margin required.

As far as sensitivity to noise is concerned, generally speaking, the sense-amplifier response depends on the precise phase and frequency of the relevant transients. As a result, very specific transient analysis is required. Often, static analysis is used as a conservative means to characterize noise response, and it has been observed that for certain circuit structures (e.g. bit-cells), static analysis captures the worst-case conditions [138][40] since the possibility of sustained noise is considered. In general, unequal phase-delays to differencing nodes in the amplifier can lead to worse conditions than those assumed by a static analysis; however, reset biasing in the NSR-SA is meant to enforce balanced conditions throughout the gain path, and, as a result, margins to sustained noise are characterized with the intention of gaining a conservative estimate. As an example, Figure 4-13 shows how transient spikes on V_{DD} can affect the critical nodes (X/Y) of the NSR-SA without causing output errors (Figure 4-13b); but sustained noise steps cause stringent noise conditions that are more likely to cause output errors on QB (Figure 4-13c).

Since the NSR-SA relies on dynamic charge-storage to set the voltage of internal nodes, a purely static analysis is impractical. Instead, the procedure shown in Figure 4-14 is used. Here, a transient simulation is performed, and during the RST phase, the nominal value is applied to each of the noisy inputs (i.e. V_{DD}, V_{SS}, BL, and SUB). However, immediately following RST, a deviation step is applied to one of



(a)



(b) (c)

Figure 4-13: NSR-SA (a) circuit showning noise sensitive nodes (X/Y), and (b) reponse of X/Y due to transient spikes on V_{DD}, and (c) Response of X/Y leading to output errors on QB due to sustained step on V_{DD}.

the noisy inputs, and the value of that deviation is incrementally swept over the simulation time. To derive a transfer characteristic with respect to deviation in the noisy input (e.g. ΔV_{DD}) , the output value (QB) is sampled 2ns after de-assertion of RST, which is the longest read access-time expected.

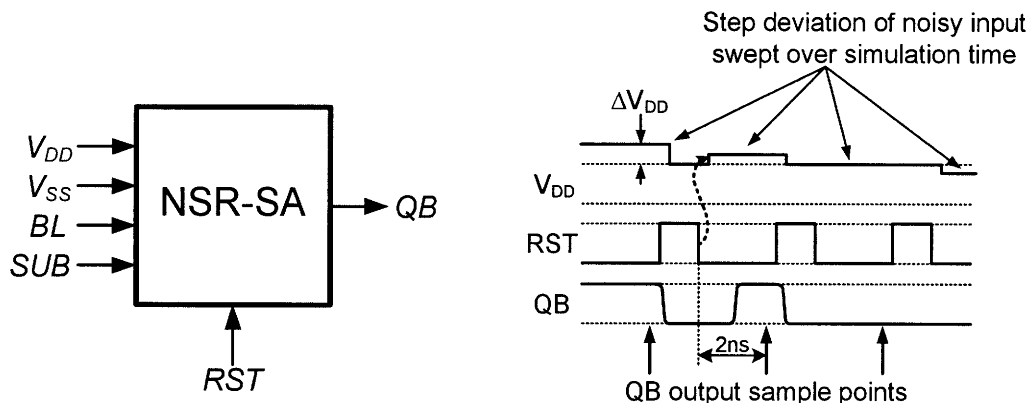


Figure 4-14: NSR-SA noise measurement simulation setup.

Using this simulation technique, the noise response of the NSR-SA is characterized below:

(1) **Bit-line noise.** In determining the desired bit-line noise margin, it is worth considering potential noise sources within the sub-array. Transitioning signals that can couple to the bit-line include the bit-line precharge control (PRE), column select control ($cSEL$), word-line (WL), and adjacent bit-lines. Proper layout (including shielding) can minimize these but cannot guarantee elimination of their impact. PRE, $cSEL$, and WL couple to the BL as shown in Figure 4-15a. Both PRE and WL cause a positive noise step on BL. Although this increases the bit-line droop required in order to cross the sense-amplifier trip-point, it opposes false regeneration and can hence be overcome. On the other hand, $cSEL$ causes the net noise step on BL to be in the negative direction (by nearly 5mV, as shown in the waveforms of Figure 4-15a), and noise margin is necessary to distinguish between this noise step and intentional bit-line signal droop.

In addition to coupling from transitioning signals, the substrate also couples

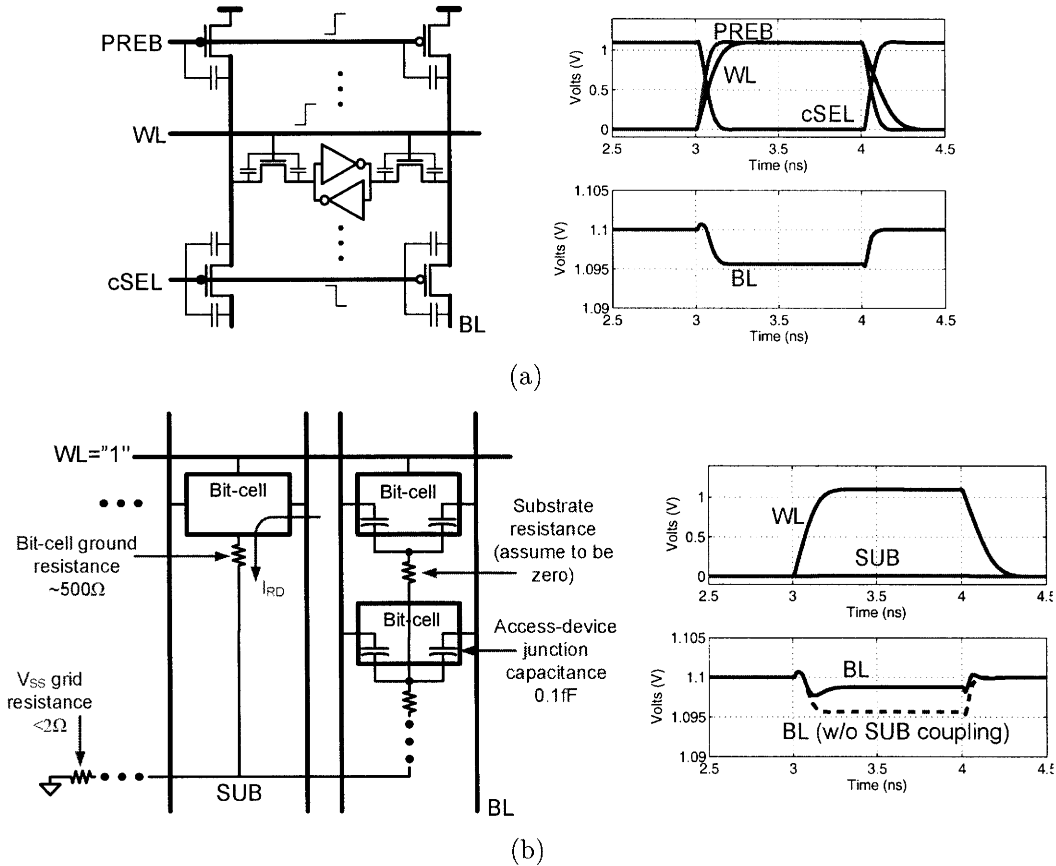


Figure 4-15: Example bit-line noise sources originating (a) from precharge, word-line, and column-select control signal coupling, and (b) substrate coupling.

to BL through the source/drain junction capacitances from the access-devices of all bit-cells sharing the bit-line. The equivalent circuit is shown in Figure 4-15b. When the word-line is asserted, the bit-cells in the accessed row sink read-current, which gets aggregated for the entire row and injected into the V_{SS} grid. The V_{SS} grid is connected to the substrate (SUB), and assuming a highly conductive substrate, the resulting voltage drop couples to the bit-lines through the junction capacitances. Figure 4-15b shows the resulting bit-line noise for a 256×256 array configuration. Here, the noise observed tends to increase the voltage of BL (as compared to the dotted waveform), and hence does not require additional noise margin. Nonetheless, as with PRE and WL coupling, the noise increases the bit-line voltage droop required to trip the sense-amplifier.

Based on the predictable and unpredictable noise sources, two features of the

NSR-SA are critical in order to balance the noise-margin and sensitivity requirements: (1) the ability to set the desired margin (i.e. input trip-point, V_{TRIP}) and (2) the stability of V_{TRIP} over PVT conditions.

Figure 4-16 shows the input bit-line transfer characteristic derived using the setup of Figure 4-14. Here, the transfer characteristic is derived in the presence of process variation (both global and local) and temperature variation (voltage is considered in detail below), and the NSR-SA is designed for a nominal input noise margin of 100mV. For comparison, the transfer characteristic of two inverter configurations are also shown in Figure 4-17: (a) considers a single stage inverter occupying the same active layout area as the NSR-SA and (b) considers a two stage cascaded inverter occupying the same layout area. Compared to these straight-forward inverter configurations, two features of the NSR-SA are important. First, regeneration in the NSR-SA leads to a very steep transfer characteristic. This greatly improves the speed with respect to bit-line droop, but also the noise margin selectivity (i.e. ambiguity); specifically, the steep output transition dramatically diminishes the impact of variation or noise in determining the trip-point of the following read-out stage. Second, the standard deviation of the NSR-SA due to local variation in the devices is 12mV, whereas those of the inverter configurations are 20mV and 26mV respectively. Additionally, the deviation of the nominal V_{TRIP} due to just global and temperature variation (i.e. on top of local variation), is 28mV in the NSR-SA, while that of the inverter configurations is 110mV and 98mV respectively. Further, the additional complexity of setting the desired nominal trip-points in the inverters is not considered. Accordingly, these results show that compared to inverter based single-ended sense-amplifiers, uncertainties in the noise-margin of NSR-SA (i.e. due to input errors) are minimized and high bit-line selectivity is achieved.

(2) **Power-supply noise.** Due to reset biasing before every detection phase, static noise on the power-supply cannot cause output errors in the NSR-SA. Transient

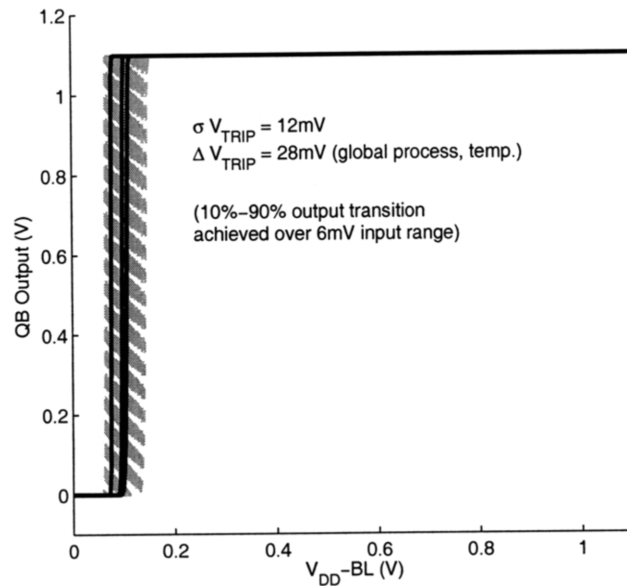


Figure 4-16: NSR-SA input transfer characteristic.

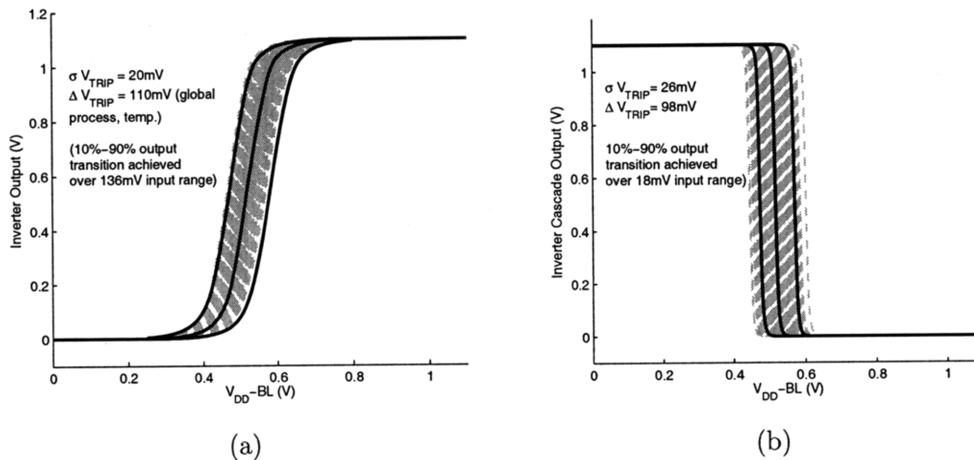


Figure 4-17: Input transfer characteristic for (a) inverter and (b) two stage inverter cascade.

noise, however, that specifically occurs after the reset phase can cause output errors. Figure 4-18 shows the transfer characteristic of such supply noise. As shown, very large negative transients are rejected at the output, as are positive transients up to 0.24V. Larger positive transients, however, do result in output errors.

Although Figure 4-18 shows that the NSR-SA is robust to output errors, the

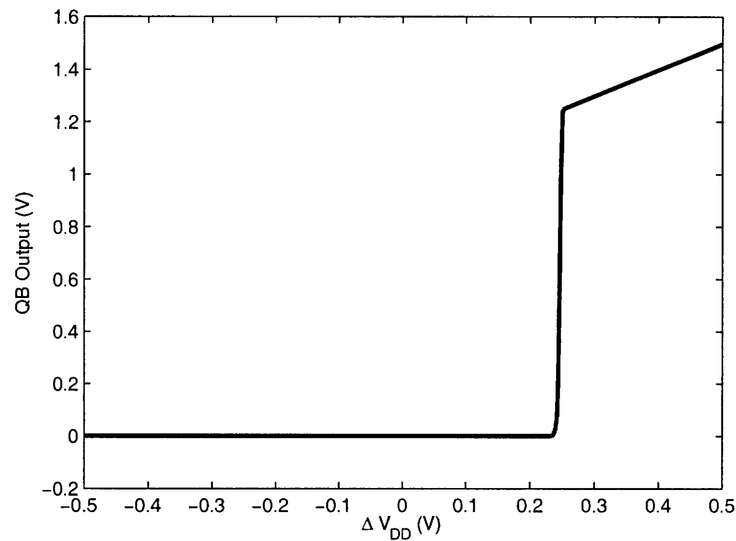


Figure 4-18: NSR-SA V_{DD} noise transfer characteristic.

effect of supply-noise on input errors must also be considered. Figure 4-19 shows how the input transfer characteristic of the NSR-SA changes as a result of ±50mV noise transients on the power-supply, applied immediately following the reset-phase. As shown, the deviation from the noise-free transfer characteristic (i.e. up to 35mV) is considerable and must be accounted for when determining the nominal input noise margin.

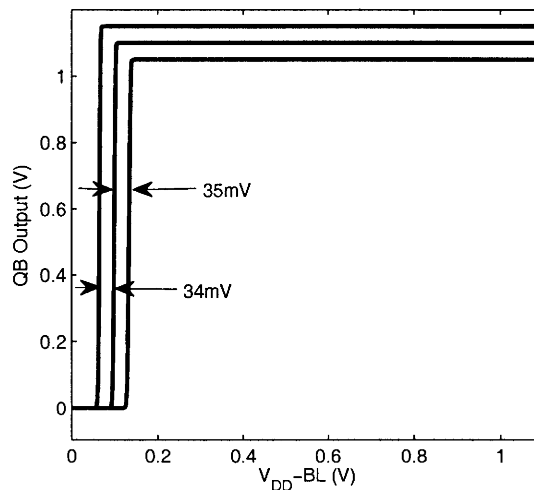
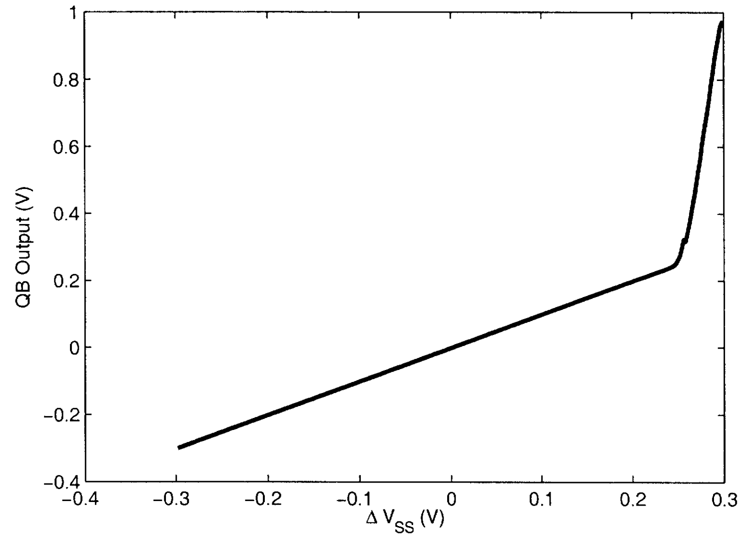
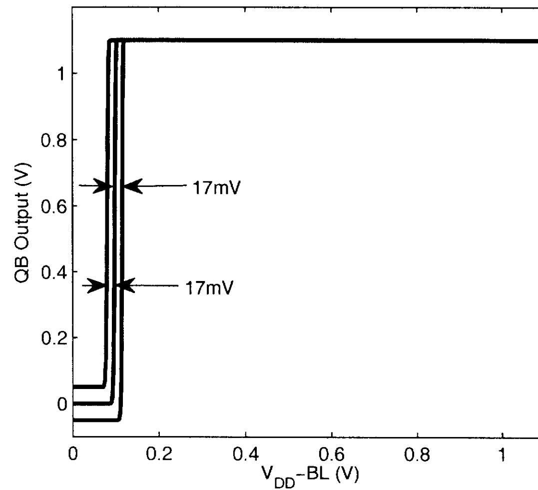


Figure 4-19: NSR-SA input transfer characteristic with ±50mV V_{DD} noise.

(3) **Ground noise.** Ground noise behaves in a complementary manner to power-supply noise. Correspondingly, the transfer characteristics illustrating both output errors and input errors are shown in Figure 4-20. Once again, the NSR-SA is highly robust to output errors, but input errors of up to 17mV are observed for ±50mV noise transients.



(a)



(b)

Figure 4-20: NSR-SA transfer characteristic for (a) V_{SS} noise and (b) input with ±50mV V_{SS} noise.

Figure 4-21 shows the input errors that result as the V_{DD} and V_{SS} noise is swept in the manner of Figure 4-14. As shown, the input trip-point can shift by 30mV

and 70mV due to 100mV of V_{SS} and V_{DD} noise respectively. Hence, these noise source must be considered when setting the nominal trip-point of the NSR-SA.

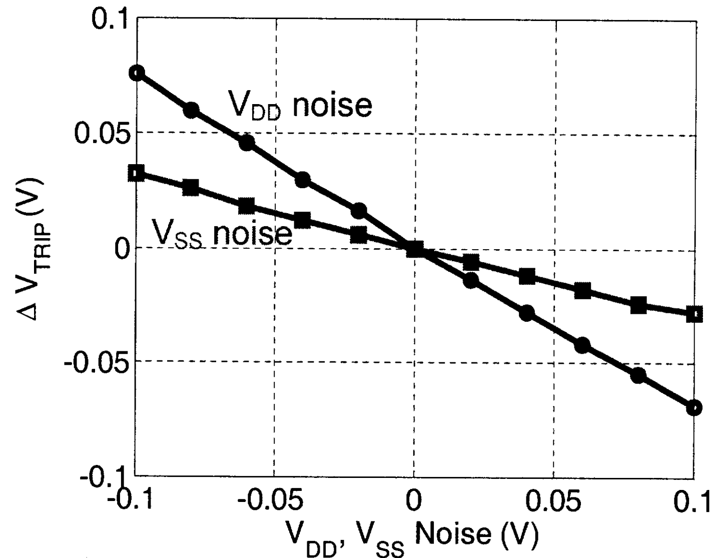


Figure 4-21: Input errors resulting from V_{DD} and V_{SS} noise.

(4) **Substrate noise.** Unlike the power-supply and ground noise, simulations show that even very large substrate noise has nearly no impact on NSR-SA output or input errors.

4.3.2 Test-Chip Architecture

A block diagram of the test-chip architecture is shown in Figure 4-22. Here, two 64kb (256 × 256) arrays of high-density $0.25\mu m^2$ 6T SRAM bit-cells are integrated. The first drives a set of conventional strobed sense-amplifiers of the structure shown in Figure 4-10, and the second drives a set of NSR-SAs, so that the relative performances can be accurately compared.

To measure the access-time with the NSR-SA, the word-line enable signal, WLE, must be asserted, and the $CLKIN$ signal must be swept inward, as shown in Figure 4-22, until bit-failures appear. To measure the access-time with the conventional sense-amplifier, however, first the $STRB$ signal must be swept inward, with the $CLKIN$ signal held at a much larger delay, until bit-errors appears. Then, to measure the

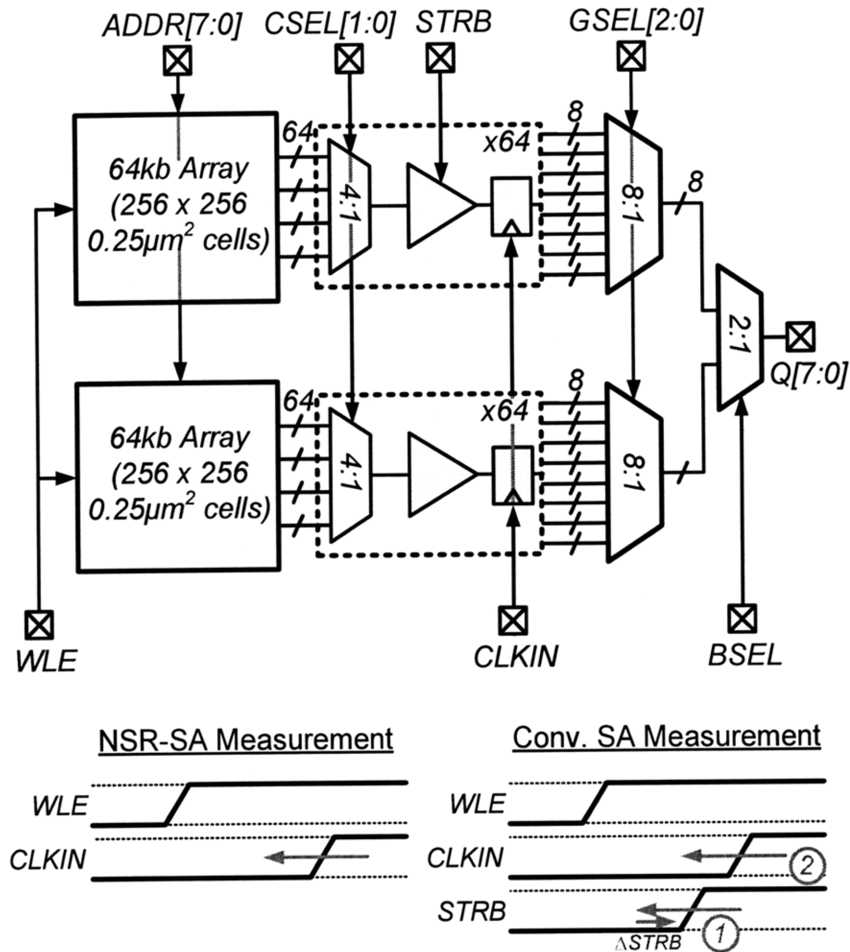


Figure 4-22: Block-diagram of prototype test-chip and access-time measurement methodology.

$WLE - CLKIN$ access-time, the $STRB$ signal must be delayed (by a predetermined amount) in order to avoid excessive sense-amplifier delays that can result from metastability. During actual measurements, this delay was conservatively set to be very large (i.e. $\Delta STRB$=1ns). Then, the $CLKIN$ signal is swept-in, and this excess $STRB$ delay is subtracted to get an estimate of the final $WLE - CLKIN$ access-time. This two sweep procedure is also shown in Figure 4-22. The final $WLE - CLKIN$ access-times can then be suitably compared, since the WLE and $CLKIN$ paths for the two arrays are carefully matched on-chip.

Bit-line Noise Margin Measurement Circuit

As mentioned in Section 4.2, with any single-ended sensing scheme it is important to characterize the noise-rejection versus sensitivity performance. In the case of bit-line noise, this requires dedicated additional circuitry, which is shown in Figure 4-23, to inject a controllable noise-amplitude on one set of bit-lines. In particular, the coupling capacitors, C_{NOISE}, can be pulsed from off-chip, and the bit-line noise can be estimated from the amplitude of the off-chip pulse (set by the potentiometer) and the nominal ratio of the capacitor divider formed by C_{NOISE} and the actual parasitic bit-line capacitance, C_{BL}. Simultaneously, the regeneration trip-point of the NSR-SA can be adjusted by injecting a controllable current at the inverter outputs using $M7$ and $M8$, whose gates are biased off-chip. This allows both the bit-line noise and NSR-SA sensitivity to be varied, allowing noise-rejection to be characterized versus access-speed.

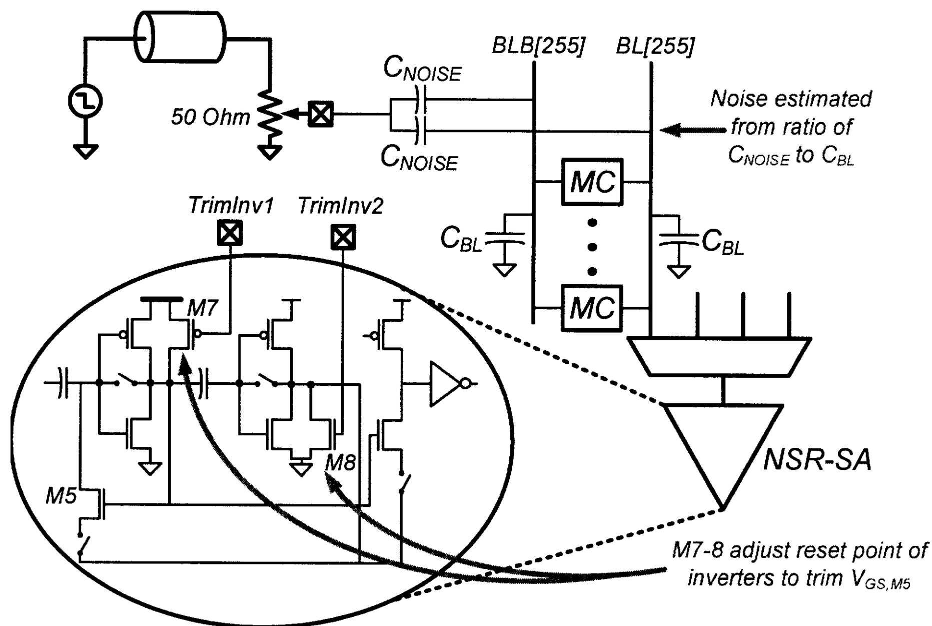


Figure 4-23: Dedicated circuitry to inject a controllable noise-amplitude on one set of bit-lines and independently adjust the sensitivity/noise-rejection of the NSR-SA.

4.3.3 Measurements and Characterization

A photograph of the prototype test-chip, which is implemented in 45nm LP CMOS, is shown in Figure 4-24. Measurements were taken from 53 chips in order to capture some statistical behavior. Two sets of test patterns were written and read from the in order to perform the characterization: (1) a checker-board pattern and its complement and (2) a binary count and its complement down the rows of the array.

With the supply-voltage set to 1V, the access-time distributions for the NSR-SA and the conventional sense-amplifier are shown in Figure 4-25. Additionally, a distribution of the difference between the access-times on each chip is also plotted to de-embed the absolute delays through board traces and chip packaging. As shown, the NSR-SA achieves superior sigma, as expected from Monte Carlo simulations. Unlike Figure 4-10, however, the measurement results include the effect of variation in the bit-cells, and as a result the absolute delays here are larger. Nonetheless, especially in the presence of this extra variation, which implies lower worst-case bit-cell read-current, the NSR-SA performs very well, achieving a factor of four reduction in access-time sigma. Accordingly, the overall worst-case delay improves from 2.46ns to 1.63ns, representing a speed-up of 34%.

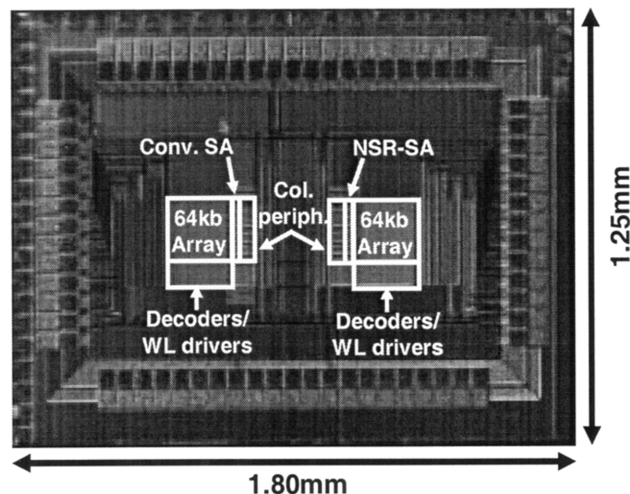


Figure 4-24: IC die photo of prototype implemented in low-power 45nm CMOS to compare performance of NSR-SA with conventional sense-amplifier.

The bit-line noise-rejection observed with respect to access-time is shown in Figure

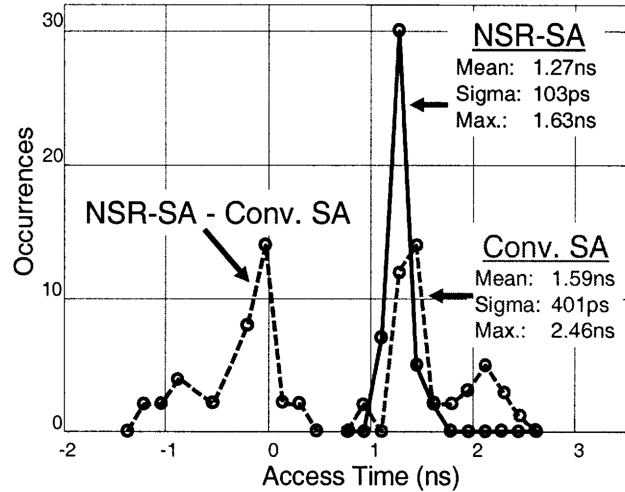


Figure 4-25: Access-time measurements from 53 chips (at 1V) showing a factor of four improvement in the NSR-SA distribution sigma compared to the conventional sense-amplifier sigma.

4-26. Here, the NSR-SA is tuned for each plotted access-time point, and the bit-line noise is increased until erroneous data is observed, yielding the corresponding bit-line noise rejection point. In the case shown, for instance, the NSR-SA can be tuned for nearly 50mV of noise-margin, corresponding to an increased amount of bit-line discharge required for data sensing. Alternatively, the speed can be increased by tuning for increased bit-line discharge sensitivity at the cost of noise-margin.

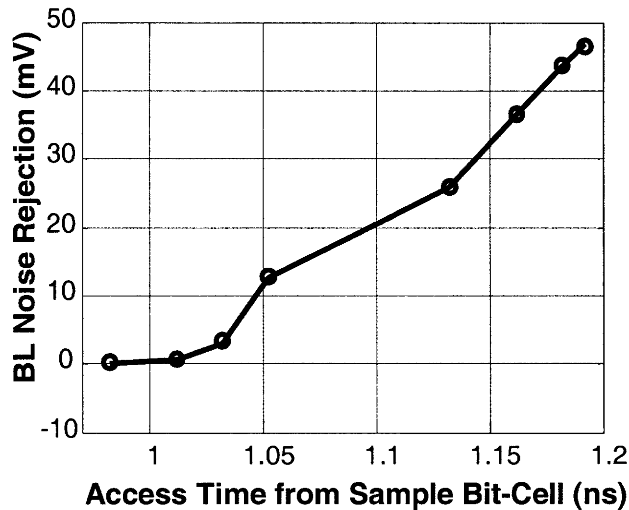


Figure 4-26: Measured bit-line noise-rejection with respect to access-time, showing ability to tune one at the cost of the other.

A performance summary of the NSR-SA and conventional sense-amplifier is provided in Table 4.1. Each conventional sense-amplifier occupies a layout area of $12\mu m^2$, while each NSR-SA occupies $19\mu m^2$, though this includes all of the testability features that have been integrated for characterizations purposes only. Additionally, during the reset phase, each NSR-SA draws static-power of 23μW while its inverters are at their trip-points; in total, the NSR-SA sense-amplifiers contribute 7% to the overall array power when the array is running at 100MHz.

Table 4.1: Test-chip performance summary.

	Conventional SA	NSR-SA
Technology	45nm low-power CMOS	
Cell size	0.25μm²	
Array configuration	256 x 256	
Capacity	64kb	
Area	12μm²	19μm² *
Power in reset	-	23μW
% of array power (at 100MHz)	2%	7%
Mean access-time	1.59ns	1.27ns
Max. access-time	2.46ns	1.67ns
Access-time sigma	401ps	103ps

*Includes testability features

4.4 Summary and Conclusions

Improving sub-array performance without increasing the supply-voltage or reducing the threshold-voltage is critical for minimizing the total energy, as lower access delays enable more aggressive supply- and threshold-voltage scaling for a given performance constraint. Performance, however, depends on bit-cell read-current as well as sense-amplifier delay and operating margins (which are associated with both timing and voltage offsets). Unfortunately, bit-cell and sense-amplifier density constraints, which are critical metrics in very high density sub-arrays, strongly oppose straight-forward techniques to improve their performance.

In the case of the bit-cell, for instance, raising the read-current and alleviating

its degradation due to variation requires strengthening the access and driver devices; in 6T cells, however, it is critical to maintain a required ratio between the driver and access device strengths for sufficient read SNM; as a result, increasing the read-current requires a large increase in cell area. 8T cells overcome the trade-off between read-current and read SNM, and offer several options for performance enhancement without area increase; however, they require specialized single-ended sensing techniques, which thus far have relied primarily on full-swing read-bit-lines. Accordingly, to retain performance, the read-bit-lines must be drastically shortened (down to eight cells [103][13]), which reduces their capacitance but degrades sub-array density.

Although sense-amplifiers face severe density and offset trade-offs themselves, their area is less constrained than the bit-cells'. Additionally, divergence between the sense-amplifier strobe delay and the read-bit-line discharge delay is emerging as a major performance limitation. Accordingly, alternate sensing structures stand to significantly improve performance with minimal impact on the total sub-array area.

A non-strobed, offset-compensating, regenerative sense-amplifier (NSR-SA) is presented in this chapter to provide small-signal, single-ended sensing capability. This makes it compatible with 8T bit-cells, and, by allowing minimal discharge on long read-bit-lines, it improves array area-efficiency, performance, and active-switching-energy. The NSR-SA's performance, compared to a standard strobed sense-amplifier, is better by up to 34%, with 4x reduction in delay sigma, when integrated with $0.25\mu m^2$ high-density bit-cells with 256 cells per read-bit-line in a 45nm LP CMOS technology.

With any single-ended sensing scheme, however, sensitivity to noise is a major concern. Accordingly, the presented sense-amplifier is analyzed for noise margins, suggesting that specific transient profiles can affect the input trip-point. Hence, the ability to set a desired read-bit-line noise margin in the sense-amplifier is a critical feature, along with stability of the noise margin in the presence of process and temperature variations.

Chapter 5

Conclusions

This work analyzes SRAMs with two objectives: (1) to determine how their energy can be minimized, and (2) to determine how their structure can be modified at the circuit and architecture levels to facilitate reliable operation under the optimal energy conditions. Energy minimization tends to require low supply-voltages and high threshold-voltages, and the main challenge associated with these is robustness in the presence of variation. Of course, device variation is also emerging as one of the primary limitations to technology scaling. As a result, ultra-low-energy techniques that primarily aim to improve robustness in the presence of extreme variation can have much broader usefulness for general circuits targeting highly advanced technologies.

5.1 Summary of Contributions

The first contribution made by this work is *SRAM energy analysis*, and two important directions for energy minimization are established from this: (1) ultra-low-voltage operation, and (2) performance improvement. Based on this, the second contribution is analysis and development of *circuits for ultra-low-voltage SRAM*, and the third contribution is analysis and development of *circuits to improve SRAM performance*. The details of these contributions are summarized below:

SRAM Energy Analysis

- Energy characterization for practical SRAM considering an idle-mode power-management strategy (which is essential for energy minimization).

- Identification of major SRAM energy components (i.e. E_{ACT}, E_{LKG}, E_{IDL}, and E_{OH}), and characterization of the parameters affecting these components.

- Isolation of V_{DD} and V_t targets for minimizing SRAM energy, and investigation of how these depend on various performance constraints. Generally, the need to aggressively reduce V_{DD} and raise V_t is motivated, for which performance constraints pose an important limitation.

- Characterization of critical SRAM metrics (i.e. read-margin, write-margin, hold-margin, and read-current) with respect to V_{DD} and V_t in the presence of variation. All of these degrade drastically at low-voltages, but read-margin and write-margin are the dominating functionality limitations.

- Characterization of the impact of variation (which itself depends on V_{DD} and V_t) on SRAM energy. Performance degradation limits V_{DD} and V_t scaling, degrading the total energy. Increased V_{DRV} degrades the energy savings achieved during idle-mode, severely impacting the absolute energy in low-performance-constraint cases.

Ultra-Low-Voltage SRAM Design

- Identification of major low-voltage SRAM functionality failures (in LP technology) arising from basic MOSFET characteristics, variation, and manufacturing defects, and analysis of the overhead imposed, using standard topologies, in order to manage these. The failures include read-margin, write-margin, and bit-line leakage.

- Analysis of 8T bit-cell, in terms of operating margins and read-current, for low-voltage. 8T bit-cell achieves better area-efficiency than 6T at ultra-low-voltages.

- Proposal of ultra-low-voltage 8T bit-cell incorporating peripheral circuit assists for aggressive voltage scaling.

- Proposal and analysis of sense-amplifier redundancy in order to improve area-offset trade-off. Despite the added selection circuitry overhead, redundancy holds promise even for highly scaled sensing networks into the 22nm node.

- Development and testing of ultra-low-voltage SRAM test-chip in 65nm LP CMOS incorporating proposed techniques and demonstrating operation down to 0.35V.

Performance Enhancement in High-Density SRAM

- Analysis of performance trade-offs and limitations in high-density SRAMs. Sense-amplifier strobe-path and array-read-path divergence is characterized and identified as a significant limitation.

- Proposal of non-strobed, regenerative sense-amplifier (NSR-SA) that provides small-signal single-ended sensing for compatibility with 8T bit-cells (which hold promise for improved energy-efficiency and performance). NSR-SA performs offset compensation to improve stability of sensing.

- NSR-SA is analyzed for input and output errors in the presence of BL, V_{DD}, V_{SS}, and substrate noise, and a technique is demonstrated to ensure desired noise-margin at the cost of sensitivity.

- NSR-SA is prototyped in 45nm LP CMOS with 64kb array of high-density $0.25\mu m^2$ bit-cells. Access-delay variability is compared against conventional strobed sense-amplifier, demonstrating 34% worst-case speed-up and 4x sigma reduction.

5.2 Concluding Thoughts and Future Directions

This work examines the key trade-offs associated with SRAM energy and how they relate to functionality and performance constraints. In all, a tight coupling exists between energy, performance, and density, and although their various interactions raise several limitations, they also increase the options for addressing target objectives. For example, performance enhancement via the sense-amplifier allows more aggressive optimization of the bit-cell for low-energy. Nonetheless, the opposing trade-offs ultimately raise the need for alternate topologies that are less constrained when it comes to energy minimization approaches. A main focus of this work is to improve the practicality of these alternate topologies by investigating and exploiting the additional dimensions of freedom they introduce. Bit-line leakage and read-current improvement via peripheral circuit-assists is an example of how the 8T bit-cell can be further enhanced in this manner. These approaches to improve the practicality and enable very aggressive application of low-energy trade-offs have led to the integration of relatively large ultra-low-energy embedded caches in demonstrated system-on-chips (SoCs) [11][23]. Finally, since the objective of density maximization is not limited just to SRAMs and, in fact, applies broadly to the semiconductor industry, limitations that emerge when the energy-density trade-off is aggressively stressed serve as a lead indication of broader scaling challenges. For instance, perhaps the much wider use of redundancy, as in the case of sense-amplifiers at ultra-low-voltages, will find increased applicability.

Upon identifying operational targets and objectives, this work aims to solve the major challenges of achieving reliable SRAM functionality under the associated conditions. Nonetheless, several major issues remain. Further, the energy trade-offs illustrated are trends that result due to some fundamental dependencies. Though these trends are difficult to overcome, the precise trade-offs governing them can be improved significantly through a variety of supplemental techniques. Here, these areas for future work are discussed briefly.

Chapter 3 and Chapter 4 present circuit techniques to expand the supply-voltage

and threshold-voltage region where the functionality and performance constraints can be met in order to facilitate energy optimization of the SRAM sub-array. Due to their dependence on variation, highlighted in Chapter 2, these constraints are statistical in nature. As a result, redundancy, combined with sophisticated repair and replacement algorithms already play an important role in overcoming SRAM failures. Increasing statistical failure sources emphasize the need for optimized error-correction techniques and simulation methodologies to accurately characterize their impact [141].

An unavoidable drawback to these approaches, however, is the overhead they introduce, both in terms of the area occupied by redundant elements and the latency imposed by error detection/correction computations. Although the area overhead must be incurred, the analysis of sense-amplifier redundancy seems to indicate that they hold very significant promise, especially in the face of rapidly increasing error-rates. To mitigate the latency overhead (as suffered with error detection and correction techniques), a variety of options must be investigated. Namely, raised supply-voltages can be heavily leveraged for ECC read-out blocks [142] since the associated digital computations are performed outside the bit-cell array, and as a result V_{DD}/V_t scaling is less urgent. Further, Other techniques for energy reduction in digital logic can also be used much more amenably [4].

Importantly, as the analysis in Chapter 2 shows, for low-performance but highly energy constrained applications, minimizing the idle-mode energy is extremely critical. Strategies that use run-time monitors to accurately estimate the hold-margin have an important role so that the idle-mode V_{DRV} can be aggressively minimized. However, approaches to actually improve the hold-margin to further reduce V_{DRV} will also be critical. for instance, both static repair/replacement and run-time error detection/correction, have been used very effectively to alleviate read-failures; but in applications limited by data-retention energy, they can be used specifically to enable further V_{DRV} scaling. Of course, in addition to hold-margin enhancement, technology optimizations to reduce idle-mode leakage sources (i.e. sub-threshold current, gate-current, junction-current, etc.) are also necessary.

Depending on the relative magnitude of the active- and idle-mode energy compo-

nents, many applications might require even more aggressive leakage-power reduction. For instance, the performance constraints considered in Chapter 2 lead to a strong dependence on both the active-energy and the leakage-energies. However, many applications can require substantially longer data-retention periods, greatly elevating the relative importance of idle-mode leakage-energy. For these applications, alternate embedded memory technologies must investigated. Specifically flash memory promises ultra-low leakage-power but incurs increased switching-energy for active data writes due to the need for high voltages to program the cells. Hybrid memory architectures, utilizing small SRAM caches in conjunction with flash arrays, have been proposed to mitigate write energy [143]. Accordingly, leveraging flash memory in highly energy-constrained applications with long data-retention periods can be a powerful means to reduce the total SRAM energy. Although additional processing steps are required, TSMC offers embedded flash memory with a 180nm CMOS technology (at the cost of seven additional mask layers), and it plans to offer the technology for an LP 90nm node in the future.

The analysis of Chapter 2 considers the average energy of homogeneous sub-arrays for an architecturally flat cache. In reality, however, hierarchical caches (and register-files [144]) can reduce the total cache access energy, thanks to reduced switching overhead in small caches [145]. With such a cache structure, the energy trade-offs and sub-array constraints change, requiring new analysis for optimal targets.

As the emphasis on embedded SRAMs increases, along with the need to optimize sub-arrays for specific applications, the optimization targets on-chip will be highly diverse. This is also true in the case of composite sub-arrays of a heterogeneous hierarchical cache. Accordingly, the biasing required will be extremely demanding. As a result, very efficient and compact embedded power-supplies (as in [146]) will play an increasing role.

Of course, they will also offer the possibility of richer dynamic voltage scaling (DVS), allowing dynamic optimization of the SRAM performance and energy. As the applications for highly energy-constrained digital circuits increase, performance responsiveness through DVS will be a critical feature of low-energy SRAMs. Though

highly scalable proof-of-concepts have been demonstrated [103][88], their practicality in energy-constrained systems requires achieving near sub-threshold operation with much better array-efficiency, as the use of reconfigurability, even in the periphery, can have a severe area cost.

Bit-cells, such as the 8T topology, that are free of the read-margin limitation hold great promise for low-energy SRAMs. As a result, array-level optimizations associated with these need to be investigated much more extensively. Chapter 3 highlights the many new optimization options afforded at the bit-cell level; however, at the array-level new biasing circuit assists and layout structures can also have a major impact. An important direction for these is a structure that enables efficient column interleaving to ease soft-error ECC and allow sense-amplifier multiplexing.

Beyond this, the 8T bit-cell also provides dual port access. This feature can be highly beneficial for specific computations (such as in-place FFT algorithms), and should be investigated. In a similar manner, the 8T bit-cell can be extended to support multiple additional read-ports with minimal extra area cost. To enable simultaneous multiple address reads, all that is required is additional two-device read-buffers that are coupled to the existing bit-cell; the ability to read from multiple addresses at the same time can have a significant impact for many architectures and applications.

Of course, single-ended sensing, to further support 8T bit-cells is an important area for continued investigation. Generally, sense-amplifiers, due to their strong impact on the overall array-performance, are an important leverage-point to enhance SRAM speed and energy. The NSR-SA of Chapter 4 begins to illuminate, and even address, the major concerns associated with sense-amplifiers in general, and those associated with single-ended sensing in particular. However, it can be extended in several ways: (1) being made more amenable for ultra-low-voltage operation (currently, its reset-phase requires a supply-voltage of almost $2V_t$); (2) being made more compact through the use of highly area-efficient coupling capacitors; and (3) being integrated with careful system timing control to minimize V_{DD} and V_{SS} noise, which stands to worsen the noise margin required with any single-ended sensing approach. In principle, the predictable node-activity during sub-array accesses should help towards managing

the issues associated with supply-noise from IR drop and coupling-noise from other sources.

Finally, the primary objective of this work has been to improve the efficiency of SRAMs for highly-energy constrained applications. Since they have posed a dominating power, performance, and reliability concern in the associated systems, SRAM use has had to be very cautious. With improvements in their energy and reliability, however, it is the hope that these systems can employ architectures that preferentially favor embedded SRAM use, opening up the possibility of many more architectural options. As a result, severely energy-constrained applications will also benefit from the superior trade-offs afforded by highly-parallel and memory-rich processor architectures.

Driven by low-energy SRAMs, these architectural approaches will be applied more and more aggressively. To facilitate them, 3D integration is emerging as an important mainstream technology solution for extreme-parallelism and extreme multi-core. It is motivated primarily by an insatiable need for embedded-memory. Although, 3D integration is unlikely to improve the basic metrics of the SRAM sub-array (such as access-delay) [147], it greatly enhances the use-ability of SRAMs by introducing high-bandwidth, flexible, and low overhead access-interfaces. As a result, energy minimization techniques (like those presented in this work) will increase the appeal of SRAMs, and 3D memory-rich integration will support their heavy utilization in highly-energy constrained applications [148].

Chapter 6

Appendix A: Acronyms

six-transistor SRAM bit-cell (6T)

seven-transistor SRAM bit-cell (7T)

eight-transistor SRAM bit-cell (8T)

nine-transistor SRAM bit-cell (9T)

ten-transistor SRAM bit-cell (10T)

complementary metal-oxide semidonductor (CMOS)

drain induced barrier lowering (DIBL)

digital signal processor (DSP)

dynamic voltage scaling (DVS)

error correction coding (ECC)

general-purpose (GP)

integrated circuit (IC)

low-power (LP)

metal-oxide semiconducting field-effect transistor (MOSFET)

negative bias temperature instability (NBTI)

N-channel metal-oxide semiconductor (NMOS)

non-strobed regenerative sense amplifier (NSR-SA)

P-channel metal-oxide semiconductor (PMOS)

process, voltage, and temperature (PVT)

random dopant fluctuation (RDF)

reverse-short-channel effect (RSCE)

strong-arm flip-flop (SAFF)

static noise margin (SNM)

system-on-chip (SoC)

static random access memory (SRAM)

voltage transfer characteristic (VTC)

Bibliography

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, April 1965.

[2] ——, "No exponential is forever: But "forever" can be delayed!" in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2003, pp. 20–23.

[3] A. Khakifirooze and D. A. Antoniadis, "The future of high-performance CMOS: Trends and requirements," in *IEEE European Solid-State Device Research Conference*, Sept. 2008, pp. 30–37.

[4] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proceedings of IEEE*, vol. 83, no. 4, pp. 498–523, April 1995.

[5] M. Horowitz and W. Dally, "How scaling will change processor architecture," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 132–133.

[6] H. Yamauchi, "Embedded SRAM trend in nano-scale CMOS," in *IEEE Int. Workshop on Memory Technology, Design and Testing*, Dec. 2007, pp. 19–22.

[7] K. Zhang, F. Seigneret, H. Yamauchi, H. Pilo, H. Shiral, and M. Hatanaka, "Embedded memory design for nano-scale VLSI systems," in *IEEE Int. Solid-State Circuits Conf. Forum*, Feb. 2008.

[8] E. J. Marinissen, B. Prince, and D. K.-S. Y. Zorian, "Challenges in embedded memory design and test," in *Proc. of Design, Automation and Test in Europe Conference and Exhibition*, March 2005, pp. 722–727.

[9] V. George, S. Jahagirdar, C. Tong, K. Smits, S. Damaraju, S. Siers, V. Naydenov, T. Khondker, S.Sarkar, and P. Singh, "Penryn: 45-nm next generation intel core 2 processor," in *Proc. IEEE Asian Solid-State Circuits Conference*, Nov. 2007, pp. 14–17.

[10] ARM, "Arm1176jz(f)-s," `http://www.arm.com/products/CPUs/ARM1176.html`.

[11] J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann, and A. Chandrakasan, "A 65nm sub-v_t microcontroller with integrated SRAM and switch-capacitor DC-DC converter," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 318–319.

[12] J. Pille, C. Adams, T. Christensen, S. Cottier, S. Ehrenreich, F. Kono, D. Nelson, O. Takahashi, S. Tokito, O. Torreiter, O. Wagner, and D. Wendel, "Implementation of the CELL broadband engine in a 65nm SOI technology featuring dual-supply SRAM arrays supporting 6GHz at 1.3V," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 322–323.

[13] R. Joshi, R. Houle, K. Batson, D. Rodko, P. Patel, W. Huott, R. Franch, Y. Chan, D. Plass, S. Wilson, and P. Wang, "6.6+ GHz low Vmin, read and half select disturb-free 1.2 Mb SRAM," in *Proc. IEEE Symp. VLSI Circuits*, June 2007, pp. 250–251.

[14] H. Pilo, V. Ramadurai, G. Braceras, J. Gabric, S. Lamphier, and Y. Tan, "A 450ps access-time SRAM macro in 45nm SOI featuring a two-stage sensing-scheme and dynamic power management," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 378–379.

[15] L. Wong, S. Hossain, A. Ta, J. Edvinsson, D. Rivas, and H. Naas, "A very low-power CMOS mixed-signal IC for implantable pacemaker applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 12, pp. 2446–2456, 2004.

[16] L. Padeletti and S. S. Barold, "Digital technology for cardiac pacing," *The American Journal of Cardiology*, vol. 95, no. 4, pp. 479–482, Feb. 2005.

[17] S. Kim, N. Cho, S.-J. Song, D. Kim, K. Kim, and H.-J. Yoo, "A 0.9-V 96-μW digital hearing aid chip with heterogeneous Σ-Δ DAC," in *Proc. IEEE Symp. VLSI Circuits*, June 2006, pp. 55–56.

[18] H. Neuteboom, B. M. J. Kup, and M. Janssens, "A DSP based hearing instrument IC," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 11, pp. 1790–1806, Nov. 1997.

[19] J. Georgiou and C. Toumazou, "A 126-μW cochlear chip for a totally implantable system," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 2, pp. 430–443, Feb. 2005.

[20] K. D. Wise, D. J. Anderson, J. F. Hetke, D. R. Kipke, and K. Najafi, "Wireless implantable microsystems: High-density electronic interfaces to the nervous system," *Proceedings of IEEE*, vol. 92, no. 1, pp. 76–97, Jan. 2004.

[21] S. O'Driscoll, T. Meng, K. Shenoy, and C. Kemere, "Neurons to silicon: Implantable prosthesis processor," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2006, pp. 552–553.

[22] B. Gyselinckx, C. Van Hoof, J. Ryckaert, R. Yazicioglu, P. Fiorini, and V. Leonov, "Human++: autonomous wireless sensors for body area networks," in *Proc. IEEE Custom Integrated Circuits Conference*, 2005, pp. 13–19.

[23] D. Finchelstein, V. Sze, M. Sinangil, Y. Koken, and A. P. Chandrakasan, "A low-power 0.7-V H.264 720p video decoder," in *Proc. IEEE Asian Solid-State Circuits Conference*, Nov. 2008, pp. 173–176.

[24] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, pp. 102–114, Aug. 2002.

[25] B. H. Calhoun, D. C. Daly, N. Verma, D. F. Finchelstein, D. D. Wentzloff, A. Wang, S.-H. Cho, and A. Chandrakasan, "Design considerations for ultra-low energy wireless microsensor nodes," *IEEE Transactions on Computers*, vol. 54, no. 6, pp. 727–740, June 2005.

[26] L. Schweibert, S. Gupta, and J. Weinmann, "Research challenges in wireless networks of biomedical sensors," in *Mobile Computing and Networking*, 2001, pp. 151–165.

[27] M. Hamilton, E. Graham, P. Rundel, M. Allen, W. Kaiser, M. Hansen, and D. Estrin, "New approaches in embedded network sensing for terrestrial ecological observatories," *Environmental Engineering Science*, vol. 24, no. 2, pp. 192–204, 2007.

[28] S. N. Pakzad, G. L. Fenves, S. Kim, and D. E. Culler, "Design and implementation of scalable wireless sensor network for structural monitoring," *ASCE Journal of Infrastructure Engineering*, vol. 14, no. 1, pp. 89–101, 2008.

[29] J. Paradiso and T. Starner, "Energy scavenging for mobile and wireless electronics," *Pervasive Computing, IEEE*, vol. 4, no. 1, pp. 18–27, 2005.

[30] S. Roundy, E. Leland, J. Baker, E. Carleton, E. Reilly, E. Lai, B. Otis, J. Rabaey, P. Wright, and V. Sundararajan, "Improving power output for vibration-based energy scavengers," *Pervasive Computing, IEEE*, vol. 4, no. 1, pp. 28–36, 2005.

[31] R. Amirtharajah, S. Meninger, J. Mur-Miranda, A. Chandrakasan, and J. Lang, "A micropower programmable DSP powered using a MEMS-based vibration-to-electric energy converter," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2000, pp. 362–363, 469.

[32] V. Leonov, T. Torfs, P. Fiorini, and C. Van Hoof, "Thermoelectric converters of human warmth for self-powered wireless sensor nodes," *Sensors Journal, IEEE*, vol. 7, no. 5, pp. 650–657, 2007.

[33] O. Chevalerias, T. O'Donnell, D. Power, N. O'Donovan, G. Duffy, G. Grant, and S. O'Mathuna, "Inductive telemetry of multiple sensor modules," *Pervasive Computing, IEEE*, vol. 4, no. 1, pp. 46–52, 2005.

[34] S. Mandal and R. Sarpeshkar, "Low-power CMOS rectifier design for RFID applications," *IEEE Transactions on Circuits and Systems*, vol. 54, no. 6, pp. 1177–1188, 2007.

[35] R. Signorelli, J. Schindall, and J. Kassakian, "Nanotube enhanced ultracapacitors," in *15th International Seminar on Double Layer Capacitors and Similar Energy Storage Devices*, 2005.

[36] A. P. Chandrakasan, N. Verma, and D. Daly, "Ultralow-power electronics for biomedical applications," *Annu. Rev. Biomed. Eng*, vol. 10, pp. 247–274, Aug. 2008.

[37] K. Utsumi, E. Morifuji, M. Kanda, S. Aota, T. Yoshida, K. Honda, Y. Matsubara, S. Yamada, and F. Matsuoka, "A 65 nm low power cmos platform with $0.495\mu m^2$ SRAM for digital processing and mobile applications," in *Proc. IEEE Symp. VLSI Technology*, June 2005, pp. 216–217.

[38] A. Chatterjee, J. Yoon, S. Zhao, S. Tang, K. Sadra, S. Crank, H. Mogul, R. Aggarwal, B. Chatterjee, S. Lytle, C. T. Lin, K. D. Lee, J. Kim, L. Olsen, M. Quevedo-Lopez, K. Kirmse, G. Zhang, C. Meek, D. Aldrich, H. Mair, M. Mehrotra, L. Adam, D. Mosher, J. Yang, D. Crenshaw, B. Williams, J. Jacobs, M. Jain, J. Rosal, T. Houston, J. Wu, N. S. Nagaraj, D. Scott, S. Ashburn, and A. Tsao, "A 65 nm CMOS technology for mobile and digital signal processing applications," in *IEDM Dig. Tech. Papers*, Dec. 2004, pp. 665–668.

[39] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, H. Makino, Y. Yamagami, S. Ishikura, T. Terano, T. Oashi, K. Hashimoto, A. Sebe, G. Okazaki, K. Satomi, H. Akamatsu, and H. Shinohara, "A 45nm low-standby-power embedded SRAM with improved immunity against process and temperature variations," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 326–327.

[40] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.

[41] D. Boning and S. Nassif, *Design of High Performance Microprocessor Circuits.* Wiley IEEE Press, 2000, ch. Models of Process Variations in Device and Interconnect.

[42] M. Clinton, "Variation tolerant SRAM design techniques," in *IEEE Symp. VLSI Circuits, Short-Course*, June. 2007.

[43] A. Agarwal and S. Nassif, "The impact of random device variation on SRAM cell stability in sub-90-nm CMOS technologies," *IEEE Transactions on VLSI Systems*, vol. 16, no. 1, pp. 86–97, Jan. 2008.

[44] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal suppply and threshold scaling for sub-threshold CMOS circuits," in *Proc. IEEE Comp. Society Annual Int. Symp. VLSI*, April 2002, pp. 5–9.

[45] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. IEEE Int. Symp. Quality Electronic Design*, March 2004, pp. 55–60.

[46] U. Bhattacharya, Y. Wang, F. Hamzaoglu, Y. Ng, L. Wei, Z. Chen, J. Rohlman, I. Young, and K. Zhang, "45nm SRAM technology development and technology lead vehicle," *Intel Technology Journal*, vol. 12, no. 2, pp. 110–121, June 2008.

[47] K. Noda, K. Takeda, K. Matsui, S. Ito, S. Masuoka, H. Kawamoto, N. Ikezawa, Y. Aimoto, N. Nakamura, T. Iwasaki, H. Toyoshima, and T. Horiuchi, "An ultrahigh-density high-speed loadless four-transistor SRAM macro with twisted bitline architecture and triple-well shield," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 3, pp. 510–515, March 2001.

[48] B. S. Amrutur, *Design and Analysis of Fast Low-Power SRAMs.* Thesis, Stanford University, 1999.

[49] R. W. Mann, W. W. Abadeer, M. J. Breitwisch, O. Bula, J. S. Brown, B. C. Colwill, P. E. Cottrell, W. G. Crocco, S. S. Furkay, M. J. Hauser, T. B. Hook, D. Hoyniak, J. M. Johnson, C. H. Lam, R. D. Mih, J. Rivard, A. Moriwaki, E. Phipps, C. S. Putnam, B. A. Rainey, J. J. Toomey, and M. I. Younus, "Ultralow-power SRAM technology," *IBM Journal of Research and Development*, vol. 47, no. 5/6, pp. 553–566, Sept./Nov. 2003.

[50] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2005, pp. 20–25.

[51] A. Wang and A. P. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[52] K. Kouichi, M. Takayuki, M. Kyeong-Sik, and S. Takayasu, "Two orders of magnitude leakage power reduction of low voltage SRAMs by row-by-row dynamic VDD control (RRDV) scheme," in *Proc. IEEE Int. ASIC/SoC Conference*, Sept. 2002, pp. 381–385.

[53] A. Bhavnagarwala, S. V. Kosonocky, S. P. Kowalczyk, R. V. Joshi, Y. H. Chan, U. Srinivasan, and J. K. Wadhwa, "A transregional CMOS SRAM with single logic VDD and dynamic power rails," in *Proc. IEEE Symp. VLSI Circuits*, June 2004, pp. 292–293.

[54] N. Kim, K. Flautner, D. Blaauw, and T.Mudge, "Circuit and microarchitectural techniques for reducing cache leakage power," *IEEE Transactions on VLSI Systems*, vol. 12, no. 2, pp. 167–184, Feb. 2004.

[55] H. Yamauchi, T. Iwata, H. Akamatsu, and A. Matsuzawa, "Ba 0.8 V/100 MHz/sub-5 mW-operated mega-bit SRAM cell architecture with charge-recycle offset-source driving (OSD) scheme," in *Proc. IEEE Symp. VLSI Circuits*, June 1996, pp. 126–127.

[56] K. Osada, Y. Saitoh, E. Ibe, and K. Ishibashi, "16.7-fA/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray induced multierrors," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 11, pp. 1952–1957, Nov. 2003.

[57] K. Zhang, U. Bhattachalya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A SRAM design on 65nm cmos technology with integrated leakage reduction scheme," in *Proc. IEEE Symp. VLSI Circuits*, June 2004, pp. 294–295.

[58] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, and H. Makino, "A 90 nm dual-port SRAM with 2.04 μm 2 8T-thin cell using dynamically-controlled column bias scheme," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 508–509.

[59] A. Agarwal, H. Li, and K. Roy, "A single-Vt low-leakage gated-ground cache for deep submicron," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, Feb. 2003.

[60] A. Bhavnagarwala, A. Kapoor, and J. Meindl, "Dynamic-threshold CMOS SRAM cells for fast, portable applications," in *Proc. IEEE Int. ASIC/SoC Conference*, Sept. 2000, pp. 359–363.

[61] H. Kawaguchi, Y. Itaka, and T. Sakurai, "Dynamic cut-off scheme for low-voltage SRAMs," in *Proc. IEEE Symp. VLSI Circuits*, June 1998, pp. 140–141.

[62] J. Wang and B. Calhoun, "Canary replica feedback for near-DRV standby VDD scaling in a 90 nm SRAM," in *Proc. IEEE Custom Integrated Circuits Conference*, Sept. 2007, pp. 29–32.

[63] Y. Wang, H. Ahn, U. Bhattacharya, T. Coan, F. Hamzaoglu, W. Hafez, C.-H. Jan, R. Kolar, S. Kulkarni, J. Lin, Y. Ng, I. Post, L. Wel, Y. Zhang, K. Zhang, and M. Bohr, "A 1.1GHz 12μA/Mb-leakage SRAM design in 65nm ultra-low-power CMOS with integrated leakage reduction for mobile applications," in

IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers, Feb. 2007, pp. 324–325.

[64] F. Hamzaoglu, K. Zhang, Y. Wang, H. Ahn, U. Bhattacharya, Z. Chen, Y.-G. Ng, A. Pavlov, K. Smits, and M. Bohr, "A 153Mb-SRAM design with dynamic stability enhancement and leakage reduction in 45nm high-k metal-gate CMOS technology," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 376–377.

[65] H. Mair, A. Wang, G. Gammie, D. Scott, P. Royannez, S. Gururajarao, M. Chau, R. Lagerquist, L. Ho, M. Basude, N. Culp, A. Sadate, D. Wilson, F. Dahan, J. Song, B. Carlson, and U. Ko, "A 65-nm mobile multimedia applications processor with an adaptive power management scheme to compensate for variations," in *Proc. IEEE Symp. VLSI Circuits*, June 2007, pp. 224–225.

[66] K. Itoh, M. Horiguchi, and T. Kawahara, "Ultra-low voltage nano-scale embedded RAMs," in *Proc. IEEE North-East Workshop on Circuits and Systems*, June 2006, pp. 245–248.

[67] A. Pavlov and M. Sachdev, *CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies.* Springer Netherlands, 2008, ch. Soft Errors in SRAMs: Sources, Mechanisms and Mitigation Techniques.

[68] 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. IEEE Symp. VLSI Circuits*, June 2006, pp. 15–16.

[69] Y. Tsividis, *Operation and Modeling of the MOS Transistor, 2nd Edition.* Oxford University Press, 2003.

[70] N.Ickes. Thesis, Massachusetts Institute of Technology, 2008.

[71] X. Xi, M. Dunga, J. He, W. Liu, K. Cao, X. Jin, J. Ou, M. Chan, and A. Niknejad, "BSIM4.3.0 MOSFET model- user's manual," 2003.

[72] S. Ekbote, K. Benaissa, B. Obradovic, S. Liu, H. Shichijo, F. Hou, T. Blythe, T. W. Houston, S. Martin, R. Taylor, A. Singh, H. Yang, and G. Baldwin, "45nm low-power CMOS SoC technology with aggressive reduction of random variation for SRAM and analog transistors," in *Proc. IEEE Symp. VLSI Technology*, June 2008, pp. 160–161.

[73] R. Aitken, N. Dogra, D. Gandhi, S. Becker, and A. Components, "Redundancy, repair, and test features of a 90nm embedded SRAM generator," in *Proc. IEEE Int. Symp. Defect and Fault Tolerance in VLSI Systems*, Nov. 2003, pp. 467–474.

[74] K. .Itoh and R. Takemura, "Low-voltage limitations and challenges of memory-rich nano-scale CMOS LSIs," in *Proc. IEEE ESSCIRC*, Sept. 2007, pp. 68–75.

[75] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical mos transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Special Issue of the Analog Integrated Circuits and Signal Processing Journal on Low-Voltage and Low-Power Design*, vol. 8, pp. 83–114, July 1995.

[76] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyon," *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 433–449, July/Sept. 2006.

[77] K. Takeuchi, T. Fukai, T. Tsunomura, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto, "Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies," in *IEDM Dig. Tech. Papers*, Dec. 2007, pp. 467–470.

[78] S. V. Kosonocky, A. Bhavnagarwala, and L. Chang, "Scalability options for future sram memories," in *Proc. IEEE Int. Conf. on Solid-State and Integrated Circuit Technology*, Oct. 2006, pp. 689–692.

[79] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, April 2001.

[80] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake, "Redefinition of write margin for next-generation SRAM and write-margin monitoring circuit," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2006, pp. 630–631.

[81] B. Calhoun, *Low Energy Digital Circuit Design Using Sub-Threshold Operation.* Thesis, Massachusetts Institute of Technology, 2005.

[82] A. Singhee and R. Rutenbar, "Statistical blockade: a novel method for very fast monte carlo simulation of rare circuit events, and it applications," in *Proc. of Design, Automation and Test in Europe Conference and Exhibition*, April 2007, pp. 1–6.

[83] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2008, pp. 322–329.

[84] N. Verma, J. Kwong, and A. Chandraksan, "Nanometer MOSFET variation in minimum energy subthreshold circuits," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 163–174, Jan. 2008.

[85] J. Choi and H. Cha, "Memory-aware dynamic voltage scaling for multimedia applications," *IEE Proc. Computers and Digital Techniques*, vol. 153, no. 2, pp. 130–136, March 2006.

[86] T. D. Burd and R. W. Brodersen, "Design issues for dynamic voltage scaling," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2000, pp. 9–14.

[87] B. Calhoun and A. Chandrakasan, "Ultra-dynamic voltage scaling using subthreshold operation and local voltage dithering in 90nm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2005, pp. 300–301.

[88] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 65nm SRAM achieving voltage scalability from 0.25-1.2V and performance scalability from 20kHz-200MHz," in *Proc. IEEE ESSCIRC*, Sept. 2008, pp. 282–285.

[89] V. Sze, R. Blazquez, M. Bhardwaj, and A. Chandrakasan, "An energy efficient sub-threshold baseband processor architecture for pulsed ultra-wideband communications," in *IEEE Int. Conference on Acoustics, Speech and Signal Processing*, May 2006, pp. 908–911.

[90] T. Mizumo, J.-I. Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuations using an 8k MOSFET's array," in *Proc. IEEE Symp. VLSI Technology*, May 1993, pp. 41–42.

[91] S.-W. Sun and P. G. Y. Tsui, "Limitations of CMOS supply-voltage scaling by MOSFET threshold-voltage variation," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 947–949, Aug. 1995.

[92] 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 Journal of Solid-State Circuits*, vol. 42, no. 4, pp. 820–829, April. 2007.

[93] L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, "Stable SRAM cell design for the 32nm node and beyond," in *Proc. IEEE Symp. VLSI Circuits*, June 2005, pp. 128–129.

[94] 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. W. 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," in *IEDM Dig. Tech. Papers*, Dec. 2005, pp. 671–674.

[95] R. Rodriguez, J. H. Stathis, B. P. Linder, S. Kowalczyk, C. T. Chuang, R. V. Joshi, G. Northrop, K. Bernstein, A. J. Bhavnagarwala, and S. Lombardo, "The impact of gate-oxide breakdown on SRAM stability," *IEEE Electron Device Letters*, vol. 23, no. 9, pp. 559–561, Sept. 2002.

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

[97] A. Kawasumi, T. Yabe, Y. Takeyama, O. Hirabayashi, K. Kushida, A. Tohata, T. Sasaki, A. Katayama, G. Fukano, Y. Fujimura, and N. Otsuka, "A single-power-supply 0.7V 1GHz 45nm SRAM with an asymetrical unit-β-ratio memory cell," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 382–383.

[98] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A read-static-noise-margin-free SRAM cell for low-V_{DD} and high-speed applications," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2005, pp. 478–479.

[99] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A 3-GHz 70MB SRAM in 65nm CMOS technology with integrated column-based dynamic power supply," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2005, pp. 474–475.

[100] M. Khare, S. Ku, R. Donaton, S. Greco, C. Brodsky, X. Chen, A. Chou, R. Della-Guardia, S. Deshpande, B. Doris, S. Fung, A. Gabor, M. Gribelyuk, S. Holmes, F. Jamin, W. Lai, W. Lee, Y. Li, P. McFarland, R. Mo, S. Mittl, S. Narasimha, D. Nielsen, R. Purtell, W. Rausch, S. Sankaran, J. Snare, L. Tsou, A. Vayshenker, T. Wagner, D. Wehella-Gamage, E. Wu, S. Wu, W. Yan, E. Barth, R. Ferguson, P. Gilbert, D. Schepis, A. Sekiguchi, R. Goldblatt,

J. Welser, K. Muller, and P. Agnello, "A high performance 90nm SOI technology with 0.992μm^2 6T-SRAM cell," in *IEDM Dig. Tech. Papers*, Dec. 2002, pp. 8–11.

[101] B. Calhoun and A. Chandrakasan, "A 256kb sub-threshold SRAM in 65nm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2006, pp. 480–481.

[102] T.-H. Kim, J. Liu, J. Kean, and C. H. Kim, "A high-density subthreshold SRAM with data-independant bitline leakage and virtual ground replica scheme," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 330–331.

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

[104] P. Hazucha, T. Karnik, J. Maiz, S. Walstra, B. Bloechel, J. Tschanz, G. Dermer, S. Hareland, P. Armstrong, and S. Borkar, "Neutron soft error rate measurements in a 90-nm CMOS process and scaling trends in SRAM from 0.25-μm to 90-nm generation," in *IEDM Dig. Tech. Papers*, Dec. 2003, pp. 21.5.1–21.5.4.

[105] J. Maiz, S. Hareland, K. Zhang, and P. Armstrong, "Characterization of multibit soft error events in advanced SRAMs," in *IEDM Dig. Tech. Papers*, Dec. 2003, pp. 21.4.1–21.4.4.

[106] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy, "A 32kb 10T subthreshold SRAM array with bit-interleaving and differential read scheme in 90nm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 388–389.

[107] H. Soeleman and K. Roy, "Ultra-low power digital subthreshold logic circuits," in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2000, pp. 94–96.

[108] J. Chen, L. T. Clark, and Y. Cao, "Ultra-low voltage circuit design in the presence of variations," *IEEE circuits Devices Mag.*, vol. 21, no. 1, pp. 12–20, Jan./Feb. 2005.

[109] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 5, pp. 1778–1786, Sept. 2005.

[110] J. Kwong and A. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits," in *Proc. Int. Symp. Low Power Electronics and Design*, 2006, pp. 8–13.

[111] K. Zhang, K. Hose, V. De, and B. Senyk, "The scaling of data sensing schemes for high speed cache design in sub-0.18μm technologies," in *Proc. IEEE Symp. VLSI Circuits*, June 2000, pp. 226–227.

[112] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto, "A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 4, pp. 523–527, April 1993.

[113] R. Singh and N. Bhat, "An offset compensation technique for latch type sense amplifiers in high-speed low-power SRAMs," *IEEE Transactions on VLSI Systems*, vol. 12, no. 6, pp. 652–657, June 2004.

[114] Z. Liu and V. Kursun, "High read stability and low leakage SRAM cell based on data/bitline decoupling," in *Proc. IEEE Int. Systems on Chip Conference*, Sept. 2006, pp. 115–116.

[115] Y.Ye, S. Borkar, and V. De, "A new technique for standby leakage reduction in high-performance circuits," in *Proc. IEEE Symp. VLSI Circuits*, June 1998, pp. 40–41.

[116] B. Yu, E. Nowak, K. Noda, and H. Chenming, "Reverse short-channel effects and channel-engineering in deep-submicron MOSFETs: Modeling and optimization," in *Proc. IEEE Symp. VLSI Technology*, June 1996, pp. 162–163.

[117] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "A 0.2 V, 480 kb subthreshold SRAM with 1 k cells per bitline for ultra-low-voltage computing," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 518–529, Feb. 2008.

[118] B. H. Calhoun and A. Chandrakasan, "A 256kb 65nm sub-threshold SRAM design for ultra-low voltage operation," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, March 2007.

[119] M. Pelgrom, H. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *IEDM Dig. Tech. Papers*, Dec. 1998, pp. 915–918.

[120] P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, June 2005.

[121] M. P. Flynn, C. Donovan, and L. Sattler, "Digital calibration incorporating redundancy of flash ADCs," *IEEE Transactions on Circuits and Systems-II*, vol. 50, no. 3, pp. 205–213, May 2003.

[122] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[123] K. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," in *IEDM Dig. Tech. Papers*, Dec. 2007, pp. 471–474.

[124] Y. Morita, R. Tsuchiya, T. Ishigaki, N. Sugii, T. Ipposhi, H. Oda, Y. Inoue, K. Torii, and S. Kimura, "Smallest v_{th} variability achieved by intrinsic silicon on thin BOX (SOTB) CMOS with single metal gate," in *Proc. IEEE Symp. VLSI Technology*, June 2008, pp. 166–167.

[125] K. .Itoh, "Adaptive circuits for the 0.5-V nanoscale CMOS era," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2009, pp. 14–20.

[126] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. S. K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "Low power embedded SRAM modules with expanded margins for writing," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2005, pp. 480–481.

[127] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, Jan. 2006.

[128] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1155, July 2004.

[129] N. Wang, "On the design of MOS dynamic sense amplifiers," *IEEE Transactions on Circuits and Systems*, vol. CAS-29, no. 7, pp. 467–477, July 1982.

[130] H. Geib, W. Weber, E. Wohlrab, and L. Risch, "Experimental investigation of the minimum signal for reliabel operation of DRAM sense amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 7, pp. 1028–1035, July 1992.

[131] S. H. Hong, S. H. Kim, S. J. Kim, J.-K. Wee, and J. Y. Chung, "An offset cancellation bit-line sensing scheme for low-voltage DRAM applications," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2002, pp. 154–155.

[132] K.-L. Wong and C.-K. Yang, "Offset compensation in comparators with minimum input-referred supply noise," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 5, pp. 837–840, May 2004.

[133] M.-J. Lee, W. Dally, and C. Chiang, "Low-power area-efficient high-speed I/O circuit techniques," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 11, pp. 1591–1599, Nov. 2000.

[134] K. Sohn, N. Cho, H. Kim, K. Kim, H.-S.Mo, Y.-H. Suh, H.-G. Byun, and H.-J. Yoo, "An autonomous SRAM with on-chip sensors in a 80nm double stacked cell technology," in *Proc. IEEE Symp. VLSI Circuits*, June 2005, pp. 232–235.

[135] K. Osada, J.-U. Shin, M. Khan, Y.-D. Liou, K. Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi, "Universal-Vdd 0.65-2.0V 32kB cache using voltage-adapted timing-generation scheme and a lithographical-symmetric cell," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2001, pp. 168–169.

[136] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto, "An area-conscious low-voltage-oriented 8T-SRAM design under DVS environment," in *Proc. IEEE Symp. VLSI Circuits*, June 2007, pp. 256–257.

[137] N. Tzartzanis and W. W. Walker, "A differential current-mode sensing method for high-noise-immunity, single-ended register files," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 506–507.

[138] M. Khellah, D. Khalil, D. Somasekhar, Y. Ismail, T. Karnik, and V. De, "Effect of power supply noise on SRAM dynamic stability," in *Proc. IEEE Symp. VLSI Circuits*, June 2007, pp. 76–77.

[139] C. C. Enz and G. C. Temes, "Circuit techniques for reducing the effects of op-amp imperfections: Autozeroing, correlated double sampling, and chopper stabilization," *Proc. of the IEEE*, vol. 84, no. 11, pp. 1584–1614, Nov. 1996.

[140] J.-T. Wu and B. A. Wooley, "A 100MHz pipelined CMOS comparator," *IEEE Journal of Solid-State Circuits*, vol. 23, no. 6, pp. 1379–1385, Dec. 1988.

[141] J. Kim, M. McCartney, K. Mai, and B. Falsafi, "Modeling SRAM failure rates to enable fast, dense, low-power caches," in *IEEE Workshop on Silicon Errors in Logic-System Effects*, March 2009.

[142] T. Kawahara, "Error-correcting codes for memories," in *Tutorial, IEEE Int. Solid-State Circuits Conf.*, Feb. 2007.

[143] H.-L. Li, C.-L. Yang, and H.-W. Tseng, "Energy-aware flash memory management in virtual memory system," *IEEE Transactions on VLSI Systems*, vol. 16, no. 8, pp. 952–964, Aug. 2008.

[144] J. Balfour, W. Dally, D. Black-Scaffer, V. Parikh, and J. Park, "An energy-efficient processor architecture for embedded systems," *IEEE Computer Architecture Letters*, vol. 7, no. 1, pp. 29–32, Jan. 2008.

[145] N. Ickes, D. Finchelstein, and A. Chandrakasan, "A 10-pJ/instruction, 4-MIPS micropower DSP for sensor applications," in *Proc. IEEE Asian Solid-State Circuits Conference*, Nov. 2008, pp. 289–292.

[146] Y. Ramadass and A. Chandrakasan, "Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications," in *IEEE Power Electronics Specialists Conference*, June 2009, pp. 2353–2359.

[147] A. Zia, P. Jacob, R. P. Kraft, and J. F. McDonald, "A 3-tier, 3D FD-SOI SRAM macro," in *International Conference on Integrated Circuit Design and Technology and Tutorial*, June 2008, pp. 277–280.

[148] H. Saito, M. Nakajima, T. Okamoto, Y. Yamada, A. Ohuchi, N. Iguchi, and T. Sakamoto, "A chip-stacked memory for on-chip SRAM-rich SoCs and processors," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2009, pp. 60–61.