Low Energy Digital Circuit Design Using Sub-threshold Operation

by

Benton H. Calhoun

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

Doctor of Philosophy in Electrical and Computer Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 2005

Author ....
Department of Electrical Engineering and Computer Science
December 8, 2005

Certified by ....
Anantha P. Chandrakasan
Professor, Department of Electrical and Computer Engineering
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Low Energy Digital Circuit Design Using Sub-threshold Operation

by

Benton H. Calhoun

Submitted to the Department of Electrical Engineering and Computer Science
on December 8, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering

Abstract

Scaling of process technologies to deep sub-micron dimensions has made power management a significant concern for circuit designers. For emerging low power applications such as distributed micro-sensor networks or medical applications, low energy operation is the primary concern instead of speed, with the eventual goal of harvesting energy from the environment. Sub-threshold operation offers a promising solution for ultra-low-energy applications because it often achieves the minimum energy per operation. While initial explorations into sub-threshold circuits demonstrate its promise, sub-threshold circuit design remains in its infancy.

This thesis makes several contributions that make sub-threshold design more accessible to circuit designers. First, a model for energy consumption in sub-threshold provides an analytical solution for the optimum V_{DD} to minimize energy. Fitting this model to a generic circuit allows easy estimation of the impact of processing and environmental parameters on the minimum energy point. Second, analysis of device sizing for sub-threshold circuits shows the trade-offs between sizing for minimum energy and for minimum voltage operation. A programmable FIR filter test chip fabricated in 0.18μm bulk CMOS provides measurements to confirm the model and the sizing analysis. Third, a low-overhead method for integrating sub-threshold operation with high performance applications extends dynamic voltage scaling across orders of magnitude of frequency and provides energy scalability down to the minimum energy point. A 90nm bulk CMOS test chip confirms the range of operation for ultra-dynamic voltage scaling. Finally, sub-threshold operation is extended to memories. Analysis of traditional SRAM bitcells and architectures leads to development of a new bitcell for robust sub-threshold SRAM operation. The sub-threshold SRAM is analyzed experimentally in a 65nm bulk CMOS test chip.

Thesis Supervisor: Anantha P. Chandrakasan
Title: Professor, Department of Electrical and Computer Engineering

Acknowledgments

From my first weeks at MIT, a significant part of the advice offered to incoming graduate students focused on finding the right advisor. I was blessed to end up with the best. I am so grateful to Professor Anantha Chandrakasan for his continuous support and help over the last five years. My initial plan for graduate studies totally excluded circuit design until Anantha's 6.374 class captivated my interest. From those early days, Anantha has taught me a tremendous amount about circuits and other technical topics, but the other lessons I have learned from him are even more valuable. Anantha has shown me how to pursue research with zeal while maintaining the highest ethical standards and while remaining sensitive to the needs of the people involved. Anantha unswervingly pursues the best interests of his students, as I have experienced repeatedly. I have learned a lot from watching the way that he manages a large research group while constantly supporting and promoting the students who comprise it. Thank you, Anantha, for your amazing support and advice over the years. Your leadership has enriched my experience at MIT tremendously.

I also am grateful to the other members of my thesis committee, Professor Charlie Sodini and Professor Duane Boning. From my first meeting with them for my Research Qualifying Exam and throughout the process of developing my dissertation, they have challenged me with insightful questions, provided excellent suggestions, and helped me to improve my work. Thank you for your advice and support.

My friends and colleagues in Ananthagroup have made graduate school a true joy for me. I must first thank Margaret Flaherty, who helps to keep all of us in line and sane. Margaret has bailed me out of numerous crises (many of my own making). It seems that last minute emergencies frequently land on her desk, but she is always helpful and often heroic in her response. I cannot overstate how valuable it has been to be to have the friendship and fellowship of Alice Wang and Fred Lee. Our times talking and praying together always helped me to refocus my priorities and to remember what is truly important. I have certainly felt uplifted by your support and prayers, and I know that our friendships will be lifelong. Thank you!

In my earliest days in Ananthagroup, I benefitted from the tutelage of Manish Bhardwaj and Eugene Shih. Those guys helped me to relax a little bit and showed me the ropes of how to survive MIT. How interesting that Manish is now a brand new PhD student in the group (although I still turn to him for advice!). I am extremely grateful also to Travis Simpkins, who has been a great friend since even before I joined Ananthagroup. I am so thankful that Travis was always willing to escape with me for some "R and R" to the disc golf course, the basketball court, or the mountains. I am also grateful to Dave Wentzloff for good times on the basketball court, for answering my endless questions, and for being a good friend. I really appreciate all of the other folks who make our group such a great place to work (and relax). Thanks to Frank Honoré, Theodoros Konstantakopoulos, Raul Blazquez-Fernandez, Nathan Ickes, Alex Kern, and Yogesh Ramadass. Thanks to the Canadians: Naveen Verma, Denis Daly, Daniel Finchelstein, Payam Lajevardi, Vivienne Sze, Joyce Kwong, ... and Brian Ginsburg. Thanks to Nisha Checka, Shamik Das, Rex Min, CheeWe Ng, SeongHwan Cho, Julia Cline, Johnna Powell, Puneet Newaskar, Nigel Drego, and Piyada Phanaphat. Thanks to the newbies: Tao Pan, Taeg Sang Cho, Vikram Chandrasekhar, and Fred Chen. I have thoroughly enjoyed the technical discussions, jokes, riddles, lunch-time debates, invention marketing meetings, political discussions, and good times. I also want to thank the many other MIT students whom I have known in the context of classes, technical discussions, sports, and fun.

This thesis cites several publications for which I am indebted to my co-authors: Denis Daly, Naveen Verma, Daniel Finchelstein, Dave Wentzloff, SeongHwan Cho, and especially Alice Wang and Anantha Chandrakasan. Alice was so helpful ramping me up on her work when I decided to investigate sub-threshold circuits, and Anantha always knew the right way to present our results and where to submit. I am also grateful to Siva Narendra, Paul Peter Sotiriadis, Frank Honoré, James Kao, and Curt Schurgers (and his dog) for their collaboration on other works. I appreciate the help of Naveen, Joyce, and especially Payam for reading parts of this thesis.

I want to thank the sponsors who have supported my research during my time at MIT: the Defense Advanced Research Projects Agency (DARPA) and Texas Instruments. I am very grateful to the employees of Texas instruments who have repeatedly shared their time and knowledge with me. I have thoroughly enjoyed our collaboration, and I am very impressed with this company and its wonderful employees. I want specifically to thank Alice Wang (again!), David Scott, Andrew Marshall, Terence Breedijk, Richard White, and Dennis Buss.

Although most of the work in this dissertation got done at the lab, most of my support came from home. I cannot express enough appreciation and thanks to my wife, Mary Kathryn. She has been my strength throughout our time in Cambridge, providing me with the love and caring that I needed to sustain me. She has made our home a place of refuge for me where I can find rest. Our marriage has been amazing, and I find myself falling more in love with her each day. I am so blessed to be loved by a woman who is the excellent wife from the Proverb:

> [10] An excellent wife who can find?
> She is far more precious than jewels.
> [11] The heart of her husband trusts in her,
> and he will have no lack of gain.
> Proverbs 31:10-11 (ESV)

I cannot imagine facing life without my best friend and most cherished companion. One of the many ways that my wife amazes me is with her parenting. I learn how to be a better father by watching the patience and wisdom that she uses with our children. The Lord has blessed us with two great daughters, Anna Grace and Ruby Kate, and with a son on the way! My girls are amazing. It is such a pleasure to get home from work and to soak up their love and enthusiasm! The joys of sharing life with these girls made any frustrations from work seem insignificant in comparison. Plus, they always made sure I was up early in the morning in case I was tempted to sleep late!

Our family has had the privilege of making many wonderful friends during our time in Cambridge. I especially want to thank our friends at Christ the King Presbyterian Church. We have thoroughly enjoyed sharing in your lives for the last several years as we served one another and served the Lord together. Thanks to the Kyes, the Barnes, the Ashes, the Vickers, the Hixons, the Sondereggers, the Dykxhoorns, the

Killeens, the Russells, the Woolberts, and the Litmans. We have learned a lot from you and appreciate all of the good times. I am also grateful for the community in the Westgate apartments. We love playing with all of the kids on the playground and hanging out with so many great families.

I am very thankful for my parents, Sam and Jackie Calhoun. They have provided for me generously in every way in every step of my life. I am grateful always to have them to turn to for encouragement and advice. I am especially aware of all of their prayers on our behalf. I am also thankful to my brothers, Brian and Stephen Calhoun, for lots of conversations over the last few years. I am looking forward to spending more time together in the years to come. I also want to thank my grandfather, James L. Highsmith, Sr. (Gungy to us), who gave me my first exposure to computers. As an electrical engineer himself, he was eager to introduce me to an IBM machine like the ones that he built and sold. I still clearly remember writing my first BASIC program with Gungy giving instructions over my shoulder.

I included the following verses in the acknowledgements for my Master's thesis at MIT a few years ago:

> [28]Do you not know?
>   Have you not heard?
> The LORD is the everlasting God,
>   the Creator of the ends of the earth.
> He will not grow tired or weary,
>   and his understanding no one can fathom.
> [29]He gives strength to the weary
>   and increases the power of the weak.
> [30]Even youths grow tired and weary,
>   and young men stumble and fall;
> [31]but those who hope in the LORD
>   will renew their strength.
> They will soar on wings like eagles;
>   they will run and not grow weary,
>   they will walk and not be faint.
>
> Isaiah 40:28-31 (NIV)

These words remain a description of God's faithfulness to me up until now. Although problem sets, tape-outs, and paper deadlines certainly made me weary at times, God renewed my strength when I set my hope on him. He has made me increasingly aware during my time at MIT of his grace, which he gives freely to those who seek it. I am grateful for the way that Pastor Rick Downs at CTK Presbyterian Church emphasizes God's grace in his sermons. He is quick to point out the amazing attribute of God from this verse:

> [15] For thus says the One who is high and lifted up,
> who inhabits eternity, whose name is Holy:
> "I dwell in the high and holy place,
> and also with him who is of a contrite and lowly spirit,
> to revive the spirit of the lowly,
> and to revive the heart of the contrite."
> Isaiah 57:15 (ESV)

On those rare occasions when my heart is contrite, the truth of this verse becomes very apparent. By now, I should not be surprised at God's grace because he has lavished it on me so frequently. His propensity to dwell with the contrite of heart is also perfectly consistent with the gospel message on which I depend, expressed in John 3:16 and summarized as:

> [3] ... that Christ died for our sins in accordance with the Scriptures,
> [4] that he was buried, that he was raised on the third day in accordance
> with the Scriptures,
> [5] and that he appeared to Cephas, then to the twelve.
> 1 Corinthians 15:3-5 (ESV)

As I complete my time at MIT and move on to a new phase of life, I am eager to continue to seek to magnify God's glory and to know him more fully. His faithfulness to me and to my family even when we are not faithful is certain to continue in accordance with his promises.

> [33] Oh, the depth of the riches and wisdom and knowledge of God!
> How unsearchable are his judgments and how inscrutable his ways!
> [34] "For who has known the mind of the Lord, or who has been his
> counselor?"
> [35] "Or who has given a gift to him that he might be repaid?"
> [36] For from him and through him and to him are all things. To him be
> glory forever. Amen.
> Romans 11:33-36 (ESV)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

The Integrated Circuits (ICs) world is accustomed at this point to following Moore's Law. Gordon Moore first posed his famous relation in 1965 when he observed an annual doubling of the number of transistors on a die [13]. Since that time, the IC industry has maintained the astounding exponential trends that Moore first observed by continuing to scale process technologies. As Moore joked in his 2003 keynote address to the IEEE International Solid-State Circuits Conference,"Moore's Law" now commonly refers to any set of data that, when plotted on a semilog plot, yields a straight line. His talk showed the benefits of scaling according to Moore's Law. Semiconductor revenue and transistors shipped per year have increased exponentially as the average price per transistor has dropped off exponentially with the minimum feature size. An additional benefit of this scaling is the increased performance of circuits, as Moore demonstrated by plotting MIPS versus years; sure enough, they increase exponentially. As Moore points out, however, no exponential can continue forever. In fact, process scaling has produced a number of problems that threaten to stop these exponential trends. Perhaps most notably, both the active and leakage power of processors are increasing exponentially as scaling continues [14].

As each new technology rolls out of the fab, it offers smaller transistors. The traditional goal is to reduce the minimum feature size by 30% with each new technology. This scaling theoretically will allow gate delays to decrease by 30% and area to lower by 50%. Likewise, active power should decrease for a given circuit by 30% to 65% due to smaller transistors and lower supply voltage [15]. These gains sound straight-forward, but in reality, technology scaling has become an extremely complicated process. For example, shorter transistors with thinner gate oxide require a lower supply voltage to avoid wear-out. Lowering V_{DD} too much prevents the desired decrease in delay, so the device threshold voltage, V_T, must also decrease. Lowering V_T leads to the exponential increase in sub-threshold leakage current, which increases leakage power and affects circuits like dynamic gates and memories [15]. These types of trade-offs ultimately mean that the problems facing process and circuit engineers become more complicated with each generation. Additionally, rather than simply porting old designs, chip designers take advantage of smaller devices to throw more transistors at their designs so that, for example, overall microprocessor energy increases with time and now measures over 100W [14]. Furthermore, leakage power contributes a sizeable percentage of total power [16].

High-performance applications such as microprocessors face a crop of issues related to increased power consumption. These include temperature management, heat removal, power supply networks, power density, reliability, etc. Furthermore, although it previously provided the dominant metric for almost all designs, speed is not the ultimate metric for all modern applications. Instead, a broad class of applications are emerging for which power and more specifically energy is a fundamental problem. These low power applications do not require the blazing performance of high end processors. Low power applications include portable devices such as cell phones, Personal Digital Assistant (PDA)s, and cameras. For this class of applications, lower performance requirements coupled with the need for portability demand the introduction of more stringent power constraints. Other emerging applications such as distributed sensor networks or medical applications have low energy operation as the primary concern instead of performance. These applications must maintain long system lifetimes with severely constrained energy resources. Thus, the most critical metric for a successful system is energy per operation.

All of this attention on power consumption in circuit design has motivated a significant investigation of the optimum design for minimizing energy or power. Although most of this work assumes a performance constraint, examination of voltage scaling has shown that true minimum energy operation usually occurs in the sub-threshold region ($V_{DD} < V_T$) [17][18]. While initial explorations into sub-threshold circuits demonstrate its promise, sub-threshold circuit design is not a well-developed field. A more complete understanding of how the minimum energy solution changes in different scenarios will make sub-threshold design more attractive for generic systems. Also, since many systems cannot operate exclusively in sub-threshold, it is important to explore ways to couple sub-threshold operation with higher performance modes. Finally, previous work with sub-threshold circuits has focused on combinational logic. Full sub-threshold systems are not possible without memories that function in the sub-threshold region.

This thesis focuses on extending the state-of-the-art for sub-threshold design of digital circuits by addressing these issues. The remainder of this chapter introduces micro-sensor networks as an example application class that fits well with sub-threshold circuits, points out previous work on digital sub-threshold circuits, describes the basics of digital operation in sub-threshold, and lists the contributions of this thesis.

1.2 Micro-sensor Networks and Nodes

Since the optimum supply voltage for minimizing total digital energy often occurs in sub-threshold, severely energy-constrained applications can benefit from this circuit design approach. Micro-sensor networks and the micro-sensor nodes that comprise them are one energy-constrained class of systems that provide a compelling driver for sub-threshold circuit operation primarily because the most critical limitation for micro-sensor node cost and volume is energy [19]. A micro-sensor node refers to physical hardware that provides sensing, computation, and communication functionality. A wireless micro-sensor network consists of tens to thousands of distributed nodes that sense and process data and relay it to the end-user. In the context of sensor networks, the sensor node usually is further defined to have a small form factor and an energy-constraint that is often the node's primary limitation. Typically, this constraint is imposed by the capacity of the node's battery. For this reason, most micro-sensor nodes duty cycle, or shutdown unused components whenever possible. Although duty cycling helps to extend sensor network lifetimes, it does not remove the energy constraint placed by the battery.

This section describes the basic concepts related to micro-sensor nodes, focusing on the applications for which they are designed and the subsequent requirements for energy consumption. The breadth of applications and the energy-constrained nature of micro-sensor networks points to the need for circuits that consume minimal energy and, in some cases, that can scale in performance from ultra-low levels to top-speed operation.

The fervor of research related to micro-sensor networks in recent years attests to the variety of interesting problems faced in actually implementing and using the networks. In addition to the academic value of micro-sensor networks, there are numerous real and proposed applications for putting them to use. The following section gives a look at some of the many potential applications for micro-sensor nodes.

1.2.1 Micro-sensor Applications

Proposed applications for micro-sensor networks often seem to be as numerous as the nodes in a network. This section describes some promising areas of application for micro-sensor networks to show that their requirements fit well with the strengths of sub-threshold circuits. The applications we mention are by no mean comprehensive.

Habitat Monitoring

Habitat monitoring involves long-term data collection from a natural environment, primarily for scientific study [20]. A key constraint for this application is that the solution be unobtrusive. If the process of data collection regularly disrupts the environment under observation, then it inherently changes the validity of the data itself.

Additionally, direct human disturbance can disrupt fragile animals or habitats and, even if unintentionally, can be destructive. Micro-sensor networks offer a compelling solution to this problem. If micro-sensors have a small enough form factor, they can exist in a natural environment without causing as much disruption as direct human observation (think guy in a tree with binoculars). Also, the potential for a high node density covering a large area translates to large scale data collection with high resolution. Clearly, to satisfy the requirement for unobtrusiveness, the nodes must have long lifetimes to avoid the need for re-insertion of the network. This translates directly to the need for minimizing energy per operation.

Micro-sensor networks have been employed successfully for habitat monitoring already. For example, a micro-sensor network on Great Duck Island in Maine monitors the nesting behavior of seabirds (the Leach's Storm Petrel) [21]. Specifically, the nodes collect data on the nesting environment, such as temperature, humidity, etc. Also, temperature data from inside the nests accounts for the comings and goings of the birds themselves, since the bird's body heat registers with the sensor node. The network must last for the duration of the nesting season (months) so that researchers never have to physically visit the island during that time. Thus, the energy constraint is the most important concern for this network. Furthermore, [21] proposes that compressing data on the nodes prior to communicating data can save energy overall. This points to the need more specifically for low energy digital computation.

The PODS project at the University of Hawaii gives another example of micro-sensor network for habitat monitoring. This network monitors the habitats of endangered species of plants [22]. In addition to measuring environmental parameters, the micro-sensor nodes collect, process, and send high-resolution image data. Since these plants are very fragile, the micro-sensors allow scientists to collect data without regularly visiting the site.

Environment Observation and Forecasting System (EOFS)

Environment Observation and Forecasting Systems (EOFSs) are similar to habitat monitoring, but they cover a much larger geographic area. These proposed systems collect data that can be post-processed to produce models and ultimately to forecast certain phenomena such as weather, flooding, or pollution along beaches.

One early implementation of an EOFS systems called CORIE monitors data from the lower Columbia River in Oregon [23]. Based on information gathered at different stations, the project predicts complicated circulation and mixing processes at the river's mouth. These results have implications for the habitats of local wildlife as well as for hydropower management. The initial implementation uses a relatively small number of fixed sensoring points. This type of research provides an ideal scenario for micro-sensor networks, which could extend the resolution and quality of acquired data. Also, the broad geographic distribution of the sensors makes energy efficiency important to avoid the cost of replacing all of the nodes.

The Automated Local Evaluation in Real-Time (ALERT) is a second example of an EOFS [24]. The National Weather Service implemented this system in the 1970's to monitor for potential flooding [24]. The current implementation uses a variety of different hardware. Again, this type of system would be streamlined by the use of a micro-sensor network.

Health

There are many opportunities in the category of health to use micro-sensor networks. Since many of these applications involve direct contact of the micro-sensor nodes with human patients, either implanted or external, the micro-sensor nodes must be very safe and reliable. For certain implantable nodes, additional constraints may apply, such as a limit on the amount of heat that can be dissipated due to power consumption (e.g. in the eye). Also, the energy constraint becomes very important especially for implantable nodes, because replacing a node or its battery requires surgery.

One example of an implantable micro-sensor network is an array of micro-sensors used as an artificial retina [25]. The package containing this array rests on the retina of the patient and stimulates the retina using electrical signals that physically contact the retina through micro bumps. As this type of technology progresses, it could conceivably help blind people to sense light with enough resolution to 'see'.

Another potential application is for glucose monitoring [25]. Implantable sensors could detect the glucose levels in the blood of a diabetic patient. This allows the patient to forgo daily finger pricks and provides earlier notification when insulin is needed. Other example applications for medical sensors include monitoring organs that are awaiting transplant, early detection of cancer cells via implantable sensors, and general health monitors (heart rate, blood pressure, temperature, etc.) [25].

There certainly are other opportunities to extend the lifetime of implantable medical devices (e.g. pacemakers, cochlear implants) using sub-threshold operation, but these devices are not micro-sensors.

Structural Monitoring

Structural Health Monitoring (SHM) has the goal of identifying damage in large structures [26]. A great variety of techniques fall into this category, and sensor arrays play a big part. For example, an array of sensors crafted into an optic fiber is used to monitor bridges in [27]. Likewise, sensor arrays (although not necessarily wireless arrays) are under investigation for deployment in vehicles such as aircraft [28] for monitoring structural integrity. Most SHM analysis is based on detecting the response of the structure to either ambient or applied vibrations. Wireless micro-sensors are well-suited to SHM, because they add flexibility to sensor placement and avoid the cost of running wires to numerous sensors [26]. Micro-sensor networks in this scenario also add fault tolerance and increase the rate of detection relative to other systems [29].

Other Applications

The examples of micro-sensor applications in the preceeding sections only give a glimpse of the many possibilities. Other examples include sensors in tires to monitor tire pressure and networks distributed through automotive exhaust system for emission measurement [25]. The military is interested in wireless micro-sensors for target tracking on the battlefield and for early biological and chemical weapons detection. Micro-sensor networks in cities can monitor traffic density, and networks in buildings can oversee climate control. Even Hollywood is getting excited; embedding micro-sensors in a scene during filming allows easy post-processing for synchronizing special effects and computer generated characters with the video [30]. Numerous other applications are also under investigation or being imagined [31][32].

Conservation and efficient use of energy is a key requirement for micro-sensor nodes in all of the applications that we have described. This problem must be addressed in all modes of operation. Since sub-threshold circuits can minimize energy per operation, they are an ideal choice for implementing micro-sensor nodes.

1.2.2 Micro-sensor Nodes

Over the last decade, several academic and industrial research groups have been actively designing wireless micro-sensor nodes. Most of these nodes contain sensing capability, processing or computational capability, and structures for wireless communication. Micro-sensor nodes (for low end applications) use special architectures (e.g. [33][34]). Many algorithms are developed specifically for micro-sensor nodes (e.g. distributed local algorithms [32]).

Examples of micro-sensor nodes include the Wireless Integrated Network Sensors (WINS) at UCLA [19], Crossbow Technology's Mica2 sensor node [35], the Telos mote [36], UC Berkeley's Smartdust [37], and the Massachusetts Institute of Technology's μAMPS node [38].

The earlier implementations of many of these nodes used commercial, off-the-shelf (COTS) components. Now many of the nodes use custom ICs that are tailored to the micro-sensor application space. As the nodes become more specialized, sub-threshold circuit design can make a significant impact on reducing overall power consumption in the nodes, where processor [35] power can be 50% of total power on average [33].

1.3 Energy Harvesting

For some micro-sensor applications, a limited lifetime is sufficient, and a non-rechargeable battery power is the logical choice. A 1cm^3 Lithium battery can continuously supply 10μW of power for five years [39]. However, some applications demand higher peak power or a longer lifetime in an environment where changing batteries is impractical or impossible. These types of applications require a renewable energy source. The concept of energy harvesting or energy scavenging involves converting ambient energy from the environment into electrical energy to power circuits or to recharge a battery.

Research into energy scavenging suggests that micro-sensors can utilize energy harvested from the environment. The most familiar sources of ambient energy include solar power, thermal gradients, radio-frequency (RF), and mechanical vibration. Table 1.1 [44] gives a comparison of some energy harvesting technologies. Power per area is reported because the thickness of these devices is typically dominated by the other two dimensions. The power available from these sources is highly dependent on the nodes environment at any given time. However, these examples show that it is reasonable to expect 10's of microwatts of power to be harvested from ambient energy. Thus, researchers agree that micro-sensor nodes must keep average power consumption in the 10-100μW range to enable energy scavenging [37][33]. Coupling energy harvesting hardware with the micro-sensor node can lead to self-powered micro-sensor networks (e.g. [45]). Barring significant advances in energy scavenging technology, the high instantaneous power consumption of an active wireless transceiver (mil-

Table 1.1: Examples of power densities for potential energy harvesting mechanisms

Technology	Power Density (μW/cm^2)
Vibration - electromagnetic [40]	4.0
Vibration - piezoelectric [39]	500
Vibration - electrostatic [41]	3.8
Thermoelectric (5^{o}C difference) [42]	60
Solar - direct sunlight [43]	3700
Solar - indoor [43]	3.2

liwatts for Mbps) requires micro-sensors to retain local energy storage. Coupling energy-harvesting techniques with some form of energy storage can theoretically extend micro-sensor node lifetimes indefinitely. Clearly, this type of system will greatly benefit from the significant power and energy savings made possible by sub-threshold operation.

1.4 Sub-threshold Operation

Sub-threshold circuits use a supply voltage, V_{DD}, that is less than the threshold voltage, V_T, of the transistors. In this regime, sub-threshold leakage currents charge and discharge load capacitances, limiting performance but giving significant energy savings over nominal V_{DD} operation. This section quantifies the potential savings that sub-threshold operation can achieve, gives a brief history of sub-threshold operation in digital circuits, and explains how digital circuits function in sub-threshold.

1.4.1 Motivation - Energy and Power Savings

Decreasing the supply voltage lowers both leakage power and active energy. Leakage power is simply:

$$P_{LEAK} = V_{DD} I_{LEAK} \tag{1.1}$$

Lowering V_{DD} immediately provides linear savings for leakage power based on (1.1). Table 1.2 shows what these savings might be for a range of possible scenarios.

Table 1.2: Power savings from voltage reduction alone (constant current).

	V_{DD} **nominal (V)**			
V_{DD} sub-threshold (V)	1.0	1.1	1.2	1.8
0.4	2.5X	2.75X	3X	4.5X
0.3	3.3X	3.7X	4X	6X
0.2	5X	5.5X	6X	9X

The nominal supply voltage for technologies from 65nm to 0.13μm is in the range of 1.0V to 1.2V. For 0.18μm technologies, the nominal V_{DD} is 1.8V. Depending on the technology and circuit, reasonable operating voltages for sub-threshold are in the 0.2V to 0.4V range. The table shows that leakage power decreases by from 2.5X to 9X just from the reduction in voltage.

In reality, the I_{LEAK} term from (1.1) also reduces with voltage because of Drain-Induced Barrier Lowering (DIBL). We explain the DIBL effect in more detail in Section 2.1.2. Figure 1-1 shows the reduction in I_{LEAK} that results from DIBL in

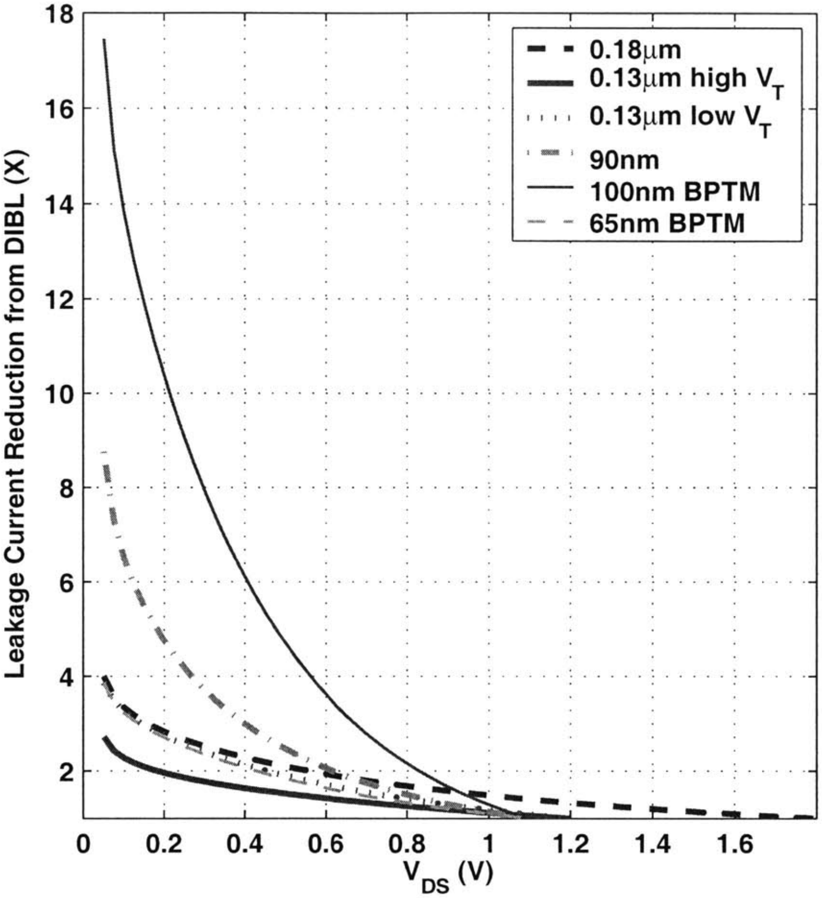


Figure 1-1: Reduction in leakage current from DIBL for different technologies. The Berkeley Predictive Technology Models (BPTMs) are from [1][2].

various technologies. If V_{DD} scales to the 0.2V to 0.4V range for sub-threshold operation, the figure shows that DIBL gives a reduction in current from 1.5X to 10X depending on the technology. Combining these savings with the savings from Table 1.2 as in (1.1) gives a possible range of savings for leakage power from $\sim$ 4X to $\sim$ 90X.

The active energy consumed by a digital circuit for an operation is the well-known:

$$E_{ACTIVE} = C_{eff} V_{DD}^2 \tag{1.2}$$

Since, to first order, the effective switched capacitance, C_{eff}, does not change, voltage scaling produces quadratic savings in active energy. Table 1.3 shows the range of possible savings over the different V_{DD} possibilities. Active energy decreases by from $\sim$ 6X to $\sim$ 80X by scaling into sub-threshold.

Table 1.3: Active energy savings from voltage reduction.

	V_{DD} **nominal (V)**			
V_{DD} sub-threshold (V)	1.0	1.1	1.2	1.8
0.4	6.25X	7.6X	9X	20.25X
0.3	11.1X	13.4X	16X	36X
0.2	25X	30.25X	36X	81X

This simplistic analysis shows the range of potential savings for both leakage power and active energy available by scaling into the sub-threshold region. Clearly, the order of magnitude or more gains that are possible make sub-threshold design a valuable avenue to explore for low energy applications and scenarios.

1.4.2 Previous Work: History of Sub-threshold

Sub-threshold digital operation was first examined theoretically in the context of studying the limits of voltage scaling in the early 1970s [46]. An analysis of an inverter operating entirely in weak inversion showed that bi-stable operating points were not possible below roughly $4V_{th}$ [46]. An implementation of a ring oscillator was shown to work at 100mV soon thereafter [47].

While sub-threshold (weak inversion) operation became very popular for low power analog circuits [48][49], digital sub-threshold circuits were slower to catch on. Techniques described as "micropower" design that were applied to energy constrained applications such as digital watches used extremely low voltages for the technologies of the day. However, these implementations stopped just short of pushing V_{DD} below V_T (e.g. [50][51]).

Theoretical interest in sub-threshold operation for digital circuits remained. It was shown that minimum energy operation occurs in the sub-threshold region [17], and plots show a minimum energy point in sub-threshold [17][52]. Then, apparently as a means of maintaining performance, the authors suggest operation at the edge of sub-threshold by setting $V_T = V_{DD}$ [17][52]. This type of analysis burgeoned into "ultra-low-power" CMOS technologies and design. V_{DD} was argued to be the most important knob for controlling CMOS power, but the accompanying decrease in speed was preventing widespread application of aggressive V_{DD} scaling [53][54]. Parallelism was seen as one method of maintaining speed at lower supply voltages [53]. Based on this sort of reasoning, ultra-low-power circuits were designed to function at extremely low voltages. Again, rather than scaling V_{DD} below V_T, ultra-low-power designers tailored V_T to scale also so that it stayed below V_{DD}. For example, Stanford ULP CMOS uses near 0 V_T devices and body biasing to tune V_T up to around $V_{DD}/3$. The resulting circuits can operate down to below 100mV, but V_T is lower still [55].

Digital sub-threshold logic continued to surface as a theoretical approach to lowering power and energy [56]. A fresh look at the calculation of lower bounds from [46] showed that improvements in technology meant weak inversion operation could theoretically occur in the 36-80mV range [57]. Again, however, implementations of ultra-low-power circuits were designed with $V_T < V_{DD}$ [57].

By the late 1990s, the increased attention on power in digital circuits made sub-threshold digital logic more interesting. A study of logic styles specifically for sub-threshold operation poses that pseudo-nMOS logic offers a lower energy solution than static CMOS for some circuits [58]. The savings in energy result primarily from reduced delay, since the power consumption of pseudo-nMOS is higher due to static current [58]. Analysis of dynamic domino circuits for sub-threshold operation appears in [59]. Again, the primary advantage of domino circuits over static CMOS is the reduced delay, which lowers energy for very active circuit blocks.

An adaptive filter design for hearing aids that uses the pseudo-nMOS logic style is reported in [60][61]. Simulation results for this filter show that it operates in sub-threshold and uses parallelism to achieve the required performance of 22kHz at 400mV. A test chip implements an 8x8 carry save array multiplier in 0.35μm CMOS to examine sub-threshold operation [61]. The multiplier uses body biasing to create symmetry between n-type MOSFET (nMOS) and p-type MOSFET (pMOS) currents and to reduce the delay variation due to changes in temperature, etc. [61]. Other chips continued to look at inverters and ring oscillators in sub-threshold. Fabrication of inverters that show sub-threshold operation at 100mV was made possible by biasing the substrate and n-wells together to match nMOS and pMOS current in [62]. Ring oscillator voltage-controlled oscillators (VCOs) controlled by both V_{DD} and body bias are functional in sub-threshold down to 80mV in [63].

A new look at minimum energy operation in 2002 shows that the minimum energy point often occurs with V_{DD} below V_T and depends on parameters such as activity factor [18]. This analysis is confirmed experimentally in [4], which presents a sub-threshold FFT processor that operates down to 180mV but has optimum energy per operation at $\sim$ 350mV (still sub-threshold). This FFT processor is the first major digital sub-threshold demonstration and brings us to the timeframe of this thesis. Now that sub-threshold operation has proven to make sense for energy constrained systems, more analysis is required for understanding the details of using it to minimize energy and for designing more complicated sub-threshold systems.

1.4.3 Inverter Operation in Sub-threshold

In sub-threshold operation, the channel of the transistors is not inverted, so currents flow by diffusion. Equation (1.3) shows a basic expression for sub-threshold drain current [64][65][66]:

$$I_{D:sub-threshold} = I_o \exp\left(\frac{V_{GS} - V_T}{nV_{th}}\right)\left(1 - \exp\left(\frac{-V_{DS}}{V_{th}}\right)\right) \tag{1.3}$$

where I_o is the drain current when $V_{GS} = V_T$ given in (1.4) [64][66].

$$I_o = \mu_o C_{ox}\frac{W}{L}(n-1)V_{th}^2 \tag{1.4}$$

For the purposes of this thesis, we assume that total drain current in sub-threshold equals sub-threshold current. Section 2.1.2 comments on the validity of this assumption. As expected for diffusion current, (1.3) shows that I_D depends exponentially on V_{GS}. V_T is the transistor threshold voltage, n is the sub-threshold slope factor

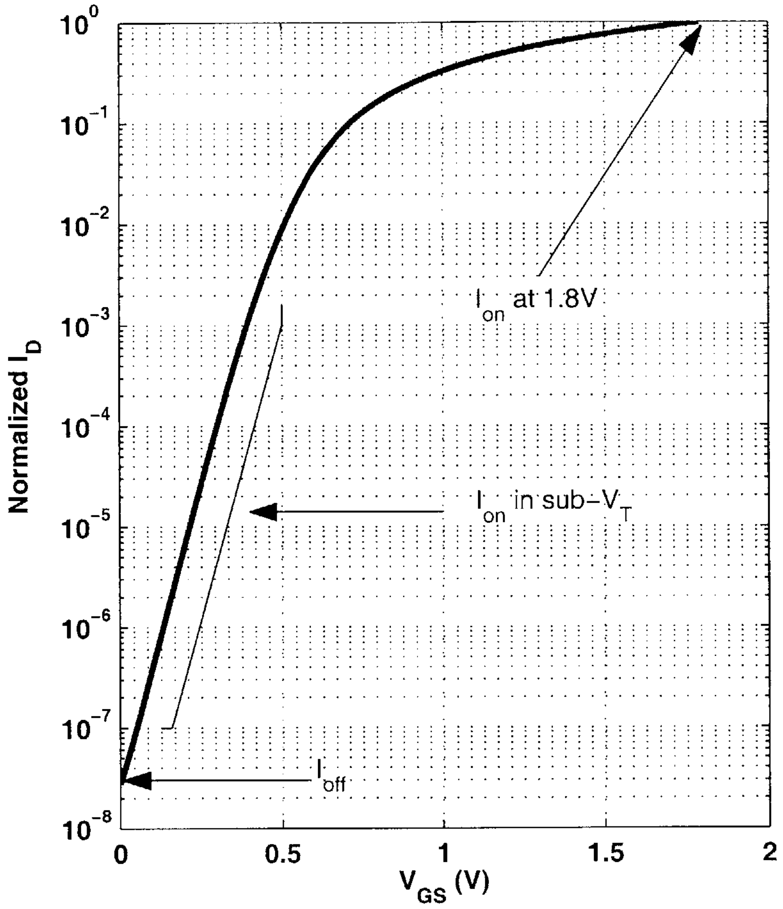


Figure 1-2: MOSFET drain current, I_D, versus gate to source voltage, V_{GS} in 0.18μm with V_{DD} = 1.8V. In sub-threshold, I_D varies exponentially with V_{GS}. In fact, we define V_T by looking at where the I_D curve deviates from its original exponential trajectory. In the sub-threshold region, the I_{on}/I_{off} ratio reduces relative to strong inversion.

($n = 1 + C_d/C_{ox}$), and V_{th} is the thermal voltage, $V_{th} = kT/q$. The parenthetical term on the right models the roll-off in current that occurs when V_{DS} drops to within a few times V_{th}.

To model Drain-Induced Barrier Lowering (DIBL), (1.3) can include a linearized DIBL coefficient, η, as in [65]:

$$I_D = I_o \exp\left(\frac{V_{GS} - V_T + \eta V_{DS}}{nV_{th}}\right)\left(1 - \exp\left(\frac{-V_{DS}}{V_{th}}\right)\right) \tag{1.5}$$

The sub-threshold slope of the transistor is defined as:

$$S = nV_{th} \ln 10 \tag{1.6}$$

S gives the inverse of the slope of I_D versus V_{GS} in millivolts per decade of change in I_D. The ideal value for S at room temperature is 60mV/decade, and it occurs at the limit when $n = 1$. Plugging (1.6) into (1.5) gives:

$$I_D = I_o 10^{\left(\frac{V_{GS} - V_T + \eta V_{DS}}{S}\right)} \left(1 - \exp\left(\frac{-V_{DS}}{V_{th}}\right)\right) \tag{1.7}$$

Figure 1-2 shows the drain current of a MOSFET versus its gate to source voltage, V_{GS} across the full range from 0V to 1.8V, which is V_{DD} for this 0.18μm technology. At low values of V_{GS} in the sub-threshold region, I_D varies exponentially with V_{GS} as expected. For the purposes of this thesis, we define the threshold voltage, V_T, by the point on the I_D versus V_{GS} plot where I_D ceases to depend exponentially on V_{GS}. This point occurs at around half a volt for the transistor in the figure.

Figure 1-2 also points out a key limitation of sub-threshold circuits. For strong inversion transistors operating at the nominal V_{DD}, the ratio of current for an *on* transistor ($V_{GS} = V_{DD}$), I_{on}, to its *off* current ($V_{GS} = 0$), I_{off}, is many orders of magnitude. In sub-threshold, however, I_{on}/I_{off} is greatly reduced. The rate at which this ratio decreases depends on S. If S is 100mV per decade, then $V_{DD} = 200mV$ implies that $I_{on}/I_{off} = 100$. The lower I_{on}/I_{off} ratio means that the circuit is slower. The *on* currents of the transistors cannot charge up the circuit's capacitances as fast. Also, this degraded ratio can lead to problems with functionality for some types of circuits [4].

Figure 1-3 shows standard I_D versus V_{DS} curves for strong inversion. The current in this plot clearly demonstrates the linear region and the velocity saturation region (not saturation, because I_D changes linearly with V_{GS} instead of quadratically). Figure 1-4 shows the same curves in sub-threshold. The curves show the exponential dependence on V_{GS}, but they otherwise appear quite similar to the strong inversion curves in their shape. The 'quasi-linear' region comes from the roll-off of current at

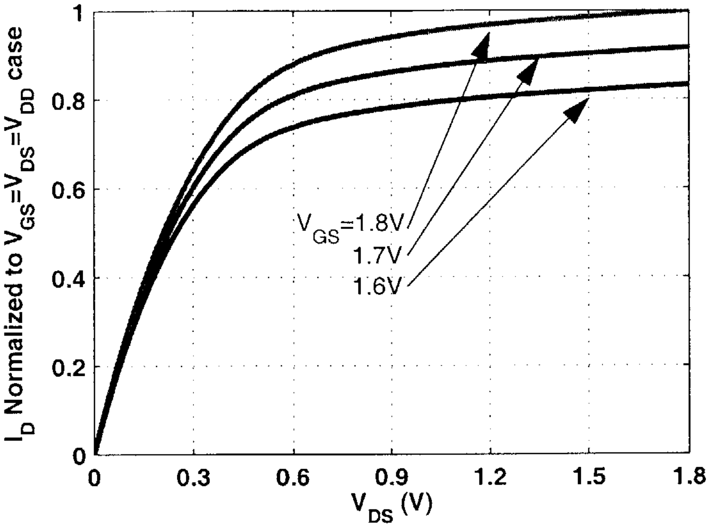


Figure 1-3: I_D versus V_{DS} curves for three values of V_{GS} in a 0.18μm process with V_{DD} = 1.8V. This above-threshold inverter shows velocity saturation, since I_D increases linearly with V_{GS}.

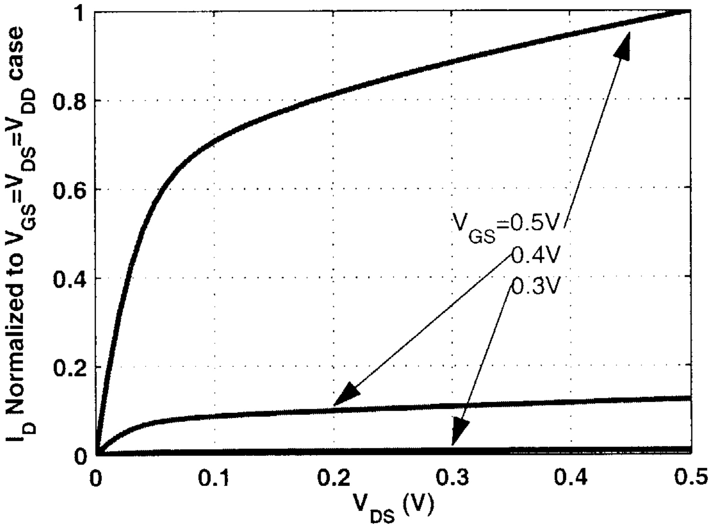


Figure 1-4: I_D versus V_{DS} curves for three values of V_{GS} in a 0.18μm process, but V_{DD} = 500mV, so the inverter is in the sub-threshold region. I_D increases exponentially with V_{GS}.

low V_{DS}, as seen in (1.3). Unlike strong inversion, the onset of this roll-off depends only on V_{DS} and not on V_{GS}. In strong inversion, the V_{DS} dependence in the velocity saturation region results from channel length modulation and is commonly modeled with the Early voltage. In sub-threshold, the V_{DS} dependence in the 'quasi-saturation' region results from DIBL and can be modeled with the DIBL coefficient as in (1.5) or (1.7).

The similarity of this curve to the strong inversion region results in familiar behavior of static logic in sub-threshold. We will look at the inverter as an example. Consider an inverter consisting of nMOS M_N and pMOS M_P. Figure 1-5 shows the I_D versus V_{DS} curves of these two transistors arranged for load line analysis in terms of the inverter's input, V_{in}, and output, V_{out}. As V_{in} goes from 0 towards $V_{DD} = 300$mV, the nMOS curves go from bottom to top on the plot, and the pMOS curves go from top to bottom. The intersections of the lines for each value of V_{in} are marked with circles. These same circles show the points on the Voltage Transfer Characteristic (VTC) that appears in Figure 1-6. Clearly, the shape of the VTC is essentially identical to above-threshold operation. The slope of the steep part of the curve depends on the sub-threshold slope of the inverters. A lower sub-threshold slope results in a steeper transition. DIBL also impacts the slope in this region. As η increases, the slope decreases.

As this VTC shows, sub-threshold digital logic gates behave very similarly to their strong inversion behavior. The most notable differences are the longer delays that result from lower on-current and the exponential dependence of I_D on V_{GS} and V_T. We provide expressions for the VTC in sub-threshold in Section 5.3.1. A good treatment of weak inversion operation for an inverter with symmetrical nMOS and pMOS appears in [67].

1.5 Thesis Contribution and Organization

Previous work on sub-threshold design has demonstrated the feasibility of sub-threshold operation, focusing primarily on logic circuits. There remains a need to expand on this work to make sub-threshold design a well-characterized and reliable option for low power systems.

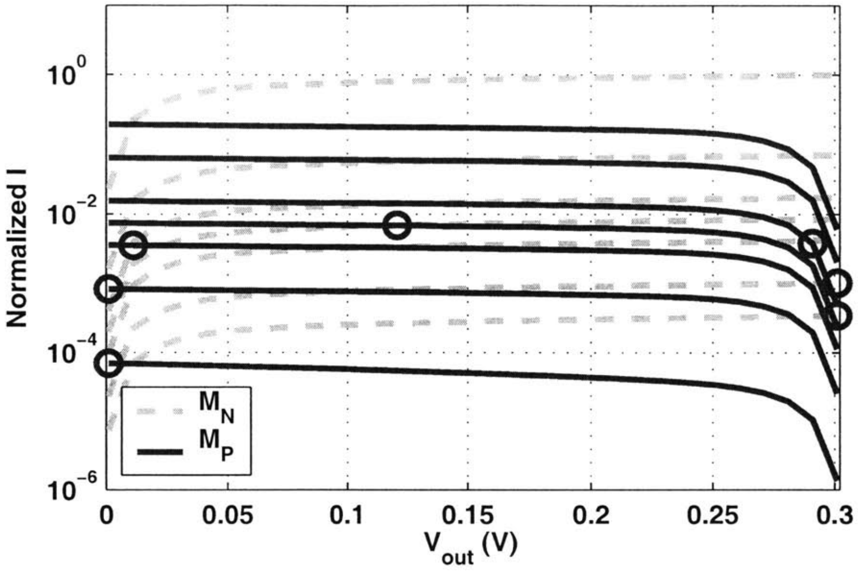


Figure 1-5: Load line analysis at the output of a sub-threshold inverter with V_{in} = [0.01, 0.05, 0.1, 0.125, 0.15, 0.2, 0.3]V.

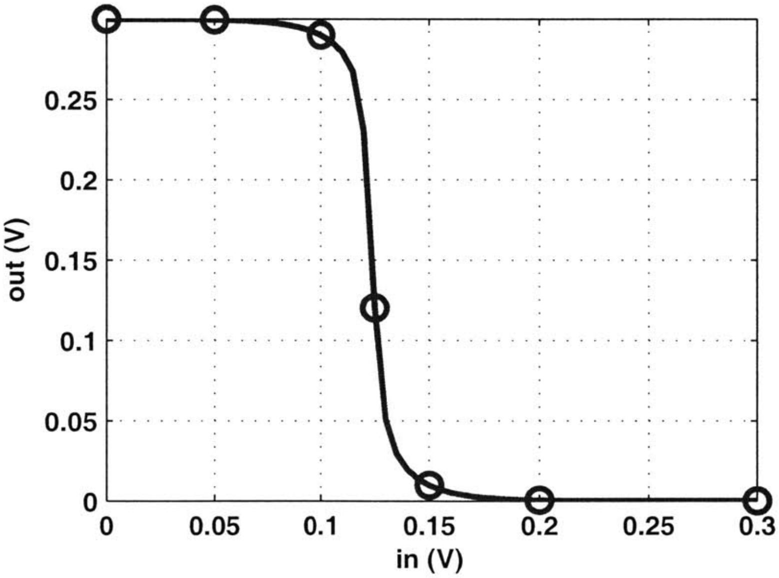


Figure 1-6: VTC for the inverter with the load lines in Figure 1-5. V_{DD} = 300mV.

The field of sub-threshold circuit design is still in its infancy. For that reason, an enormous array of potential research topics are available. Although it is impractical to address all of these topics, this thesis will investigate several key areas to advance the state-of-the-art in sub-threshold design. The broad scope covered by these areas is necessary to reveal the potential of sub-threshold operation for low power scenarios and to promote sub-threshold design as an attractive option for low power designers.

This thesis contributes in the following areas:

Modeling and Characterization

Chapter 2 provides analytical models for describing energy consumption of digital circuits in the sub-threshold region and for estimating the minimum energy point for arbitrary circuits operating in sub-threshold. Using basic equations for energy consumption, we develop an analytical solution for the optimum supply voltage to minimize total energy per operation in sub-threshold. We examine the dependencies of this solution on design and environmental parameters.

Sizing Analysis

Chapter 3 provides analysis for optimum sizing to minimize energy in sub-threshold circuits. Theoretical sizing and practical sizing for standard cell libraries are both examined. A 0.18μm sub-threshold test chip provides a comparison between sizing for minimum V_{DD} operation and standard sizing.

Ultra-Dynamic Voltage Scaling

Many applications cannot exclusively operate in sub-threshold because of the intermittent need for higher performance. Chapter 4 examines sub-threshold operation as one available mode for circuits that occasionally require high speed operation. We introduce the low overhead method of Local Voltage Dithering (LVD) for frequency and voltage scaling across a large range while minimizing energy. The strategy of Ultra-Dynamic Voltage Scaling (UDVS) combines LVD with sub-threshold operation to offer frequency and voltage scaling across the full operational range of a circuit with near-optimum energy consumption. A 90nm test chip provides measurements for the analysis.

Sub-threshold SRAM

Chapter 5 examines functionality issues for SRAM operating in sub-threshold in a 65nm process. We show that the standard six transistor SRAM bitcell and architecture cannot function below 0.6-0.7V and describe the key obstacles that lead to its failure. We introduce a bitcell designed to function in sub-threshold, and evaluate a 256kb 65nm SRAM chip that uses the bitcell.

Conclusions

Chapter 6 provides a summary of the work contained in this thesis and presents conclusions about sub-threshold circuits. It also describes some key areas for related future work.

Chapter 2

Modeling Minimum Energy Consumption

Many previous works address power minimization for a given performance constraint. In virtually every case, the performance constraint restricts operation to the strong inversion region. Early solutions for minimizing the Energy-Delay-Product (EDP) [17] are extended to account for process variation [68] and buffering options [69]. Measurements of a test chip with adaptive supply and body bias display a minimum power point for a given performance and show how forward-biased diode currents (from body biasing) can make the theoretical optimum unreachable [70]. Analytical expressions for the optimum (V_{DD},V_T) point to minimize power at a given performance are shown for transregional models based on fitted [3] or physical [71] parameters. Derivations of the sensitivities of energy and delay to different parameters support a methodology for building optimum energy circuits [72]. Taken together, these and other works give thorough attention to power optimization for strong inversion circuits. Optimizing sub-threshold circuits has received less attention.

Examining the energy-delay contours over V_{DD} and V_T shows that minimum energy operation occurs in the sub-threshold operation regime for low-to-medium performance systems, and the optimum point changes depending on activity factor and threshold variation [18]. This section develops a model to account for these effects.

2.1 Sub-threshold Current

This section reviews basic sub-threshold current equations from Section 1.4.3, introduces sub-threshold current models, and describes the impact of other components of MOSFET current in the sub-threshold region. The later sections of this chapter use the simple sub-threshold current model as a basis for minimum energy point analysis. The model that we develop for the minimum energy point allows quick estimation of important trends and the impact of key parameters on the system energy.

2.1.1 Sub-threshold Current Models

The optimization methods for strong inversion, performance-constrained optimization mentioned at the beginning of this chapter all use (2.1), which is a simplified version of (1.3), as the basic equation for modeling sub-threshold current and total off current.

$$I_{sub-threshold} = I_o \exp\left(\frac{V_{GS} - V_T}{nV_{th}}\right) \tag{2.1}$$

Some of the citations supplement this model to account for DIBL or for low V_{DS} roll-off or both, as in (1.5). We provide explanations and equations for these effects and parameters in Section 1.4.3. Other models extend the basic form of the sub-threshold current equation to account for the transitional region between weak and strong inversion. The model in [3] uses a piecewise fitted expression to smooth out the discontinuity between the regions. A complicated piecewise current model in [71] uses physical parameters to account for current across all regions of operation. The Enz, Krummenacher, and Vittoz (EKV) model has become widely used especially in the context of low-voltage analog and digital circuits [64]. The EKV model provides a single equation that simplifies to well-known expressions in weak inversion and strong inversion but that provides an accurate match for the transition region as well. The EKV model is notable for its accuracy in the transitional region as well as for the relatively small number of required parameters.

For gate level circuit design or for careful analysis of analog circuits, one of the higher order models makes sense. This is especially true when the circuit operates at the edge of weak inversion such that the simple equations no longer suffice. For the purposes of modeling the minimum energy point in this chapter, however, the simple sub-threshold current model suffices. Two reasons make this the case. First, we assume that the minimum energy point occurs well into the sub-threshold region such that (2.1) holds. Secondly, the model is intended to give an estimation of the energy-consuming behavior of large circuits, so it uses lumped fitted parameters to simplify the calculations. The subtleties of higher order current models are not necessary for this type of estimation.

One notable shortcoming of the simple sub-threshold equation is that it does not account for changes in V_T with transistor size. In Deep Sub-Micron (DSM) technologies, various higher order effects such as the Short Channel Effect, Reverse-Short Channel Effect (e.g. [73]), Narrow Channel Effect, and Reverse Narrow Channel Effect (e.g. [74]) all cause V_T, and thus I_D, to vary for small device dimensions.

2.1.2 Other Components of Current

Sub-threshold current is not the only type of leakage current that affects DSM transistors. This section briefly describes other types of current and other sub-threshold current effects and discusses their impact in the sub-threshold region.

As technologies have scaled, the gate oxide gets progressively thinner. The higher resulting electrical field across the oxide enables carriers to tunnel through the oxide [66]. None of the strong inversion energy optimization work cited above accounts for gate leakage even though gate leakage contributes significantly to total leakage in deep submicron technologies. For sub-threshold analysis, though, it is reasonable to ignore gate leakage. Gate tunneling current has a very strong dependence on the voltage across the gate, so it decreases with supply voltage much more quickly than does sub-threshold current. As a result, gate leakage becomes negligible in the sub-threshold region.

Gate-Induced Drain Leakage (GIDL) is a leakage current that arises at the edge of the drain and terminates in the body of the transistor. GIDL current appears for high V_{DS} values in combination with low V_{GS} [66]. On a semilog plot of I_D versus V_{GS}, GIDL appears as a "tail" where the current begins to increase when V_{GS} approaches 0 and continues to increase for negative V_{GS}. For sub-threshold operation, the lower V_{DS} reduces the electric field across the drain such that GIDL becomes negligible.

Section 1.4.3 has already described the effect of DIBL on sub-threshold current. DIBL occurs in short channel transistors because the depletion regions around the source and drain actually overlap slightly, lowering the source potential barrier and increasing current [66]. As V_{DS} increases, the depletion region at the drain grows, further lowering the barrier due to larger overlap with the source depletion region and increasing current. The DIBL effect still impacts current in sub-threshold, but the lower V_{DS} values mean that the magnitude of the current is lower than at larger supply voltages, lowering leakage power.

Reverse-biased diode leakage from the source and drain to the bulk also contributes to overall leakage current. This junction leakage results from a combination of minority carrier diffusion and drift at the depletion region edge and electron-hole pair generation inside the depletion region [66]. Process technologies generally are designed to make this pn-junction leakage small relative to sub-threshold current. Since the junction leakage scales with V_{DD} and temperature in a similar fashion to sub-threshold current, it generally is negligible across the full range of supply voltage.

Although these other components of current can be quite significant for normal strong inversion operation, they tend to be negligible in the sub-threshold region. One scenario where this may not be accurate is in the case of very low temperatures. Since gate leakage does not depend strongly on temperature but sub-threshold current decreases exponentially when it grows colder, gate leakage can become significant again. This is especially true if the sub-threshold current is reduced further by some other mechanism (e.g. Reverse Body Bias (RBB)). Nevertheless, except for rare cases, sub-threshold current dominates in this region of operation. This allows us to equate total current in the sub-threshold region to sub-threshold current.

2.2 Modeling the Minimum Energy Point

This section examines energy minimization for circuits operating in the sub-threshold region. We show the dependence of the optimum V_{DD} for a given technology on design characteristics and operating conditions. Solving equations for total energy provides an analytical solution for the optimum V_{DD} and V_T to minimize energy for a given frequency in sub-threshold operation. SPICE simulations of an FIR filter confirm the analytical solution and the dependence of the minimum energy operating point on important parameters.

2.2.1 Developing the Model

In this section, we derive a closed form solution for the optimum V_{DD} and V_T for a given frequency and technology operating in the sub-threshold regime ($V_{DD} < V_T$). The model we develop uses fitting parameters that are normalized to a characteristic inverter in the technology of interest. Since other gates are normalized to this inverter, its size is arbitrary. However, the minimum sized inverter is a good choice for simplicity.

Equation (2.2) shows the well-known form for the delay of an inverter in above-threshold from [3].

$$t_d = \frac{KC_gV_{DD}}{(V_{DD} - V_T)^{\alpha}} \tag{2.2}$$

The denominator of (2.2) is the *on* current of the inverter in above-threshold. We adopt the form of (2.2) for modeling the inverter in sub-threshold. Equation (2.3) shows the propagation delay of a characteristic inverter with output capacitance C_g in sub-threshold:

$$t_d = \frac{KC_gV_{DD}}{I_{o,g}\exp\left(\frac{V_{GS}-V_{T,g}}{nV_{th}}\right)} \tag{2.3}$$

As with (2.2), K is a delay fitting parameter. The expression for current in the denominator of (2.3) models the *on* current of the characteristic inverter, so it accounts for transitions through both nMOS and pMOS devices. Thus, the terms $I_{o,g}$ and $V_{T,g}$ are fitted parameters that do not correspond exactly with the MOSFET parameters of the same name unless the nMOS and pMOS are symmetrical. Operational frequency is simply:

$$f = \frac{1}{t_d L_{DP}} = \frac{1}{T_D} \tag{2.4}$$

where L_{DP} is the depth of the critical path in characteristic inverter delays and T_D is the total delay along the critical path of the circuit.

The exponential dependence of current on V_T and temperature results in large variations in delay across process and temperature corners. Likewise, the total leakage current and leakage power of a circuit will vary by a few orders of magnitude with process and temperature variation. Although this large variation is a problem for certain types of systems, it is not the focus of this chapter. In a severely energy-constrained system where the frequency of operation is less important than energy conservation, energy per operation is the more important metric. The following analysis shows that the energy per operation is much less sensitive to process and temperature changes than are delay and leakage current.

Without loss of generality, we assume that one operation occurs each cycle. We also ignore leakage from times after the current cycle because it either is accounted for in the next cycle or is addressed by a shutdown mode that is optimized separately. Dynamic (E_{DYN}), leakage (E_L), and total energy (E_T) per cycle are expressed in (2.5)-(2.7) [75][76], assuming rail-to-rail swing ($V_{GS} = V_{DD}$ for *on* current).

$$E_{DYN} = C_{eff} V_{DD}^2 \tag{2.5}$$

$$\begin{aligned} E_L &= I_{leak} V_{DD} T_D \\ &= W_{eff} I_{o,g} \exp\left(\frac{-V_{T,g}}{nV_{th}}\right) V_{DD} t_d L_{DP} \\ &= W_{eff} K C_g L_{DP} V_{DD}^2 \exp\left(-\frac{V_{DD}}{nV_{th}}\right) \end{aligned} \tag{2.6}$$

$$
\begin{aligned}
E_T &= E_{DYN} + E_L \\
&= V_{DD}^2 \left(C_{eff} + W_{eff} K C_g L_{DP} \exp\left(-\frac{V_{DD}}{nV_{th}} \right) \right)
\end{aligned} \tag{2.7}
$$

Concurrent work in [77] and [67] arrive at similar equations for sub-threshold energy consumption. Equations (2.5)-(2.7) extend the expressions for current and delay of an inverter to arbitrary larger circuits. This extension sacrifices accuracy for simplicity since the fitted parameters cannot account for all of the details of every circuit. Thus, C_{eff} is the average total switched capacitance of the entire circuit, including the average activity factor over all of its nodes. Likewise, W_{eff} estimates the average total width that contributes to leakage current. L_{DP} is the logic depth of the critical path normalized to the delay of the characteristic inverter. Solving this set of equations provides a good estimate of the optimum for the average case and shows how the optimum point depends on the major parameters. Differentiating (2.7) and equating to 0 allows us to solve for V_{DDopt}. Equation (2.8) shows the derivative of the total energy with respect to V_{DD}.

$$
\frac{\partial E_T}{\partial V_{DD}} = 2C_{eff}V_{DD} + \left(2 - \frac{V_{DD}}{nV_{th}}\right) W_{eff}KC_gL_{DP}V_{DD}\exp\left(\frac{-V_{DD}}{nV_{th}}\right) \tag{2.8}
$$

Setting (2.8) equal to zero and rearranging the equation yields:

$$
\begin{aligned}
2C_{eff}V_{DD} + \left(2 - \frac{V_{DD}}{nV_{th}}\right) W_{eff}KC_gL_{DP}V_{DD}\exp\left(\frac{-V_{DD}}{nV_{th}}\right) &= 0 \\
\left(2 - \frac{V_{DD}}{nV_{th}}\right) W_{eff}KC_gL_{DP}\exp\left(\frac{-V_{DD}}{nV_{th}}\right) &= -2C_{eff} \\
\left(2 - \frac{V_{DD}}{nV_{th}}\right) \exp\left(\frac{-V_{DD}}{nV_{th}}\right) &= \frac{-2C_{eff}}{W_{eff}KC_gL_{DP}} \\
\left(2 - \frac{V_{DD}}{nV_{th}}\right) \exp\left(2 - \frac{V_{DD}}{nV_{th}}\right) &= \frac{-2C_{eff}}{W_{eff}KC_gL_{DP}}\exp(2)
\end{aligned} \tag{2.9}
$$

The analytical solution for V_{DDopt} from (2.9) is given in (2.10):

$$
V_{DDopt} = nV_{th}\left(2 - \text{lambertW}\left(\frac{-2C_{eff}}{W_{eff}KC_gL_{DP}}\exp(2)\right)\right) \tag{2.10}
$$

If we define the argument to the Lambert W function as β:

$$\beta = \frac{-2C_{eff}}{W_{eff}KC_gL_{DP}}\exp(2) \tag{2.11}$$

then (2.10) is subject to the constraint in (2.12):

$$\beta > -\exp(-1) \tag{2.12}$$

See Appendix B regarding the Lambert W function and its constraints. Now, substituting (2.3) into (2.4) gives V_{Topt} for a given f:

$$V_{Topt} = V_{DDopt} - nV_{th}\ln\left(\frac{fKC_gL_{DP}V_{DDopt}}{I_{o,g}}\right) \tag{2.13}$$

If the argument to the natural log in (2.13) exceeds 1, then the assumption of sub-threshold operation no longer holds because $V_{Topt} < V_{DD}$. This constraint shows that there is a maximum achievable frequency for a given circuit in the sub-threshold region. Equations (2.10) and (2.13) give the optimum supply voltage and threshold voltage for sub-threshold circuits consuming the minimum energy for a given frequency. Some ultra-low power applications, such as energy scavenging sensor nodes, might consider minimizing energy to be more important than any performance requirement. Assuming a standard technology where V_T is fixed (i.e. - no triple wells for body biasing), the problem becomes finding the optimum V_{DD} and frequency to minimize energy for a given design. The optimum V_{DD} for minimizing energy per cycle in this scenario still is given by (2.10), and the optimum frequency is given by (2.4) at $V_{DD} = V_{DDopt}$.

Figure 2-1 shows the energy profile of an 8-bit, 8-tap parallel programmable FIR filter versus V_{DD}. The contributions of active and leakage energy are both shown. The lines on the plot show the results of numerical equations using a transregional current model [3], and the markers show the simulation values. The analytical solution (small star) matches the numerical model and simulations with less than 0.1% error. The optimum point is $V_{DDopt} = 250\text{mV}$ with a frequency of 30kHz. Equation (2.10)

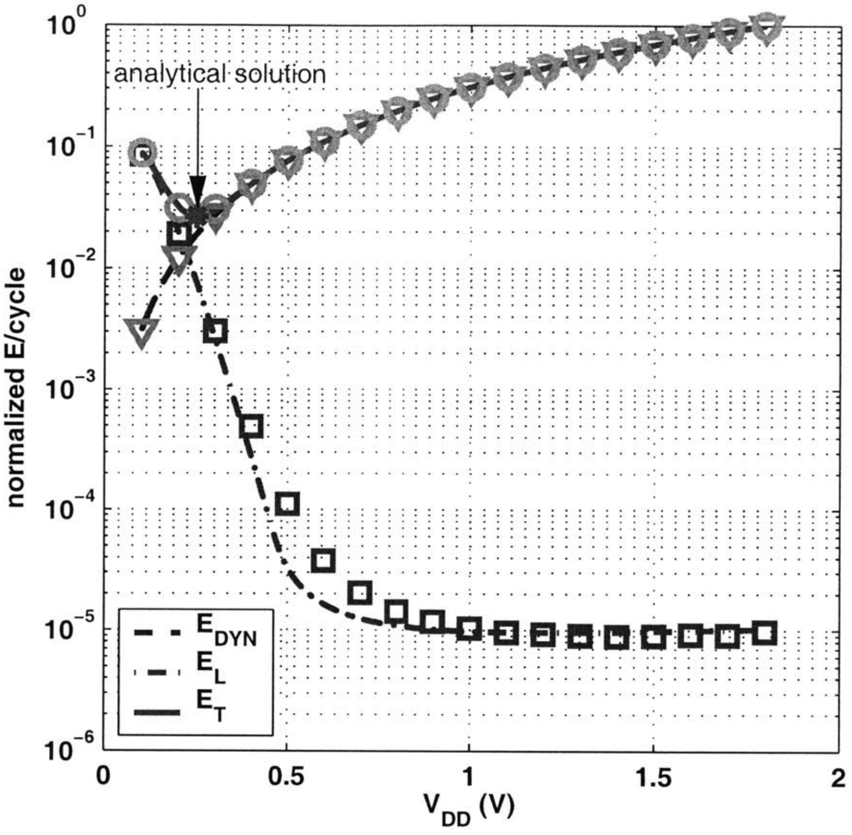


Figure 2-1: Model versus simulation of FIR filter showing minimum energy point and contribution of active and leakage energy. Markers are simulation values, lines are model [3]. Analytical solution from eqs. (2.10) and (2.7) is shown.

provided the optimum V_{DD} for the analytical solution, and substituting this value into equation (2.7) gave the total energy. Figure 2-2 shows how the delay (T_D) and current (I_{LEAK}) components of leakage energy per cycle (E_L) vary with supply voltage. As V_{DD} reduces, the current decreases due to the DIBL effect, but the delay increases exponentially in the sub-threshold region, leading to the overall increase in sub-threshold leakage energy.

Equation (2.10) shows that the optimum V_{DD} value is independent of frequency and V_T. Instead, it is set by the relative significance of dynamic and leakage energy components as expressed in equation (2.11). E_L increases compared to the characteristic inverter in two ways. First, the ratio of C_{eff}/W_{eff} decreases, indicating that a

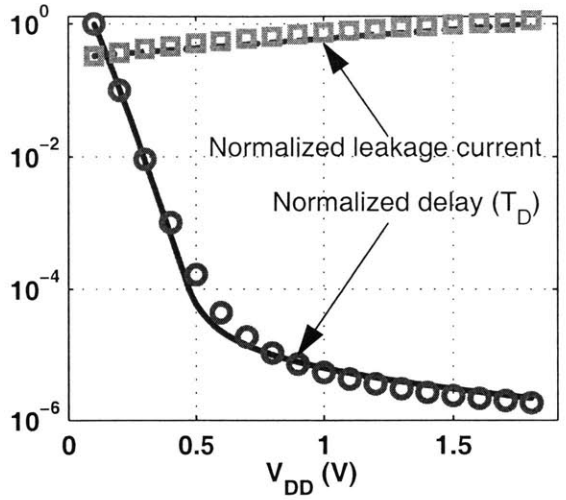


Figure 2-2: Normalized leakage current (I_{LEAK}) and delay (T_D). Markers are simulation and lines are model [3]. Although DIBL caused I_{LEAK} to decrease, the exponential increase in T_D causes leakage energy (E_L) to increase in sub-threshold.

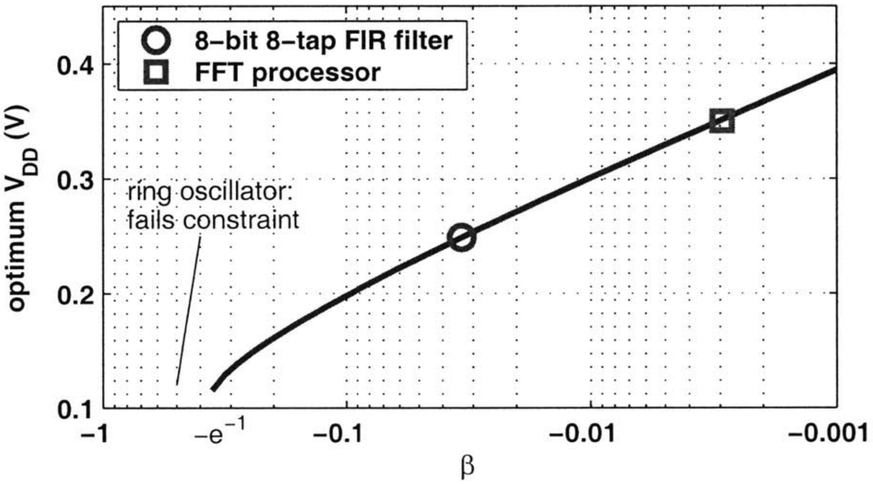


Figure 2-3: V_{DD} optimum calculated with equation (2.10). β for ring oscillator ($L_{DP} = 21$) fails constraint. β for 8x8 parallel FIR filter and scalable FFT processor [4] also shown.

greater fraction of the total width is idle and thereby drawing static current without switching. Secondly, L_{DP} can increase. The larger resulting period gives more time for leakage currents to integrate, raising E_L. Figure 2-3 shows the optimum V_{DD} values versus β for three examples. An FFT processor [4] and the FIR filter previously described have V_{DDopt} at 350mV and 250mV, respectively. The figure shows that a ring oscillator fails to meet the constraint. To see why, consider a single inverter with activity factor of one; W_{eff} and L_{DP} equal one, C_{eff} is close to C_g, and β does not meet the constraint in equation (2.12). Mathematically, this means that the derivative of E_T never equals zero. Physically, the leakage component for the single inverter with high activity remains insignificant compared to dynamic energy over all supply voltages, as shown in Figure 2-4. The true optimum V_{DD} in this case is the lowest voltage for which the circuit functions. Circuits with higher relative leakage energy, like the FIR filter or FFT processor, have less negative β and thus higher optimum V_{DD}.

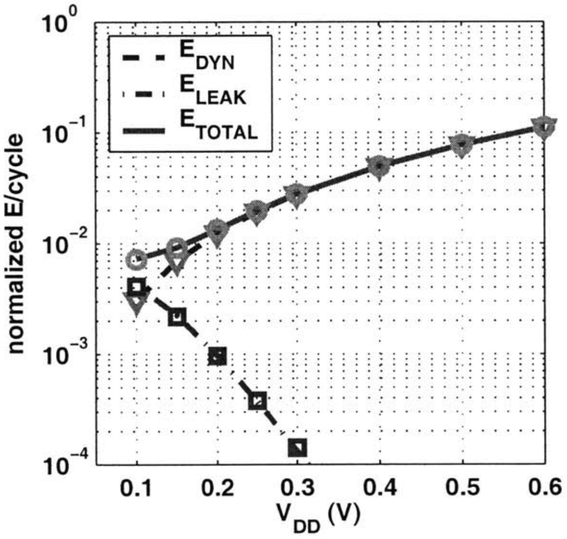


Figure 2-4: Energy per operation versus V_{DD} for a 21-stage ring oscillator has no minimum point. Markers show simulation data and lines show equations.

Figure 2-5 shows theoretically why the optimum V_{DD} is independent of V_T. As V_T decreases in the figure, the sub-threshold current increases exponentially as shown by the rise in E_L at above-V_T voltages. The sub-threshold *on* current also increases exponentially, so T_D decreases exponentially in sub-threshold and offsets the rise in I_{LEAK} such that E_L does not change in the sub-threshold region. When V_T decreases

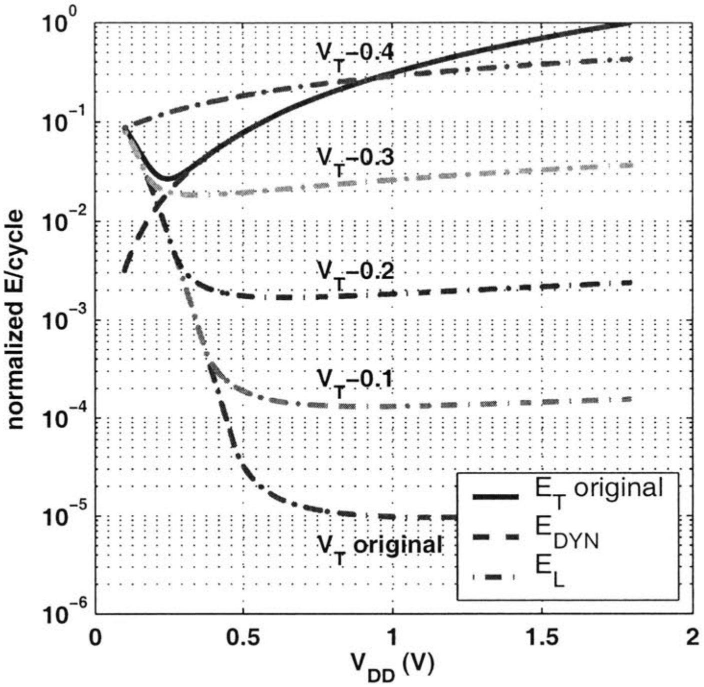


Figure 2-5: Lowering V_T does not improve the energy per operation in the sub-threshold region, but it will increase performance at the minimum energy point.

too far, then $V_{DDopt} > V_T$ so the sub-threshold equations become invalid. The figure shows that E_L physically exceeds E_{DYN} for extremely low V_T in this filter example. Of course, the advantage to lowering V_T is increased performance in the sub-threshold region for the same energy per operation.

2.2.2 Calibrating the Model

Calibrating the model to match a generic circuit requires the fitting of only three parameters once the delay and leakage of the characteristic inverter are known. C_{eff} is the average effective switched capacitance of the entire circuit, including the average activity factor over all of its nodes, short circuit current, glitching effects, etc. To calibrate the model, C_{eff} is estimated by measuring average supply current and solving $C_{eff} = I_{avg}/(fV_{DD})$. W_{eff} estimates the average total width, relative to the characteristic inverter, that contributes to leakage current. W_{eff} is determined by simulating the circuit's steady-state leakage current and normalizing to the characteristic inverter. Since W_{eff} is a function of circuit state, averaging the circuit leakage current for simulations over many states improves the total leakage estimate. Simulating to exercise the circuit's critical path, measuring its delay, and normalizing to the characteristic inverter provides the logic depth, L_{DP}. Solving this set of equations provides a good estimate of the optimum for the average case and shows how the optimum point depends on the major parameters.

2.3 Minimum Energy Point Dependencies

The optimum V_{DD} to minimize energy will vary depending on the circuit's operating scenario, environment, and temperature. This section examines the impact of these parameters on the minimum energy point.

2.3.1 Operating Scenario

Figure 2-1 graphically confirms the trend apparent in equation (2.10). Any relative increase in the leakage component of energy per cycle will push the optimum V_{DD} higher, and the frequency at the optimum point also increases. In the figure, this corresponds to any decrease in E_{DYN} or increase in E_L. Likewise, any decrease in E_L or increase in E_{DYN} will lower the optimum V_{DD}. These types of changes can occur for a given circuit without changing its intrinsic attributes.

For example, consider using the FIR filter in a system whose workload, ω, changes widely. This might be in a video context where the processing per frame depends on the difference between consecutive frames. If the current frame is nearly identical to the previous, then very little work is required. A scene change, on the other hand, could demand the maximum number of computations. Assuming the clock is gated when no computation is required (i.e. no mid-cycle shutdown is available) and normalizing to one cycle, C_{eff} per cycle becomes ωC_{eff} in equation (2.11).

Figure 2-6 shows the impact of a changing workload on the energy characteristics of the 8-bit, 8-tap FIR filter as an example. We vary the workload over three orders of

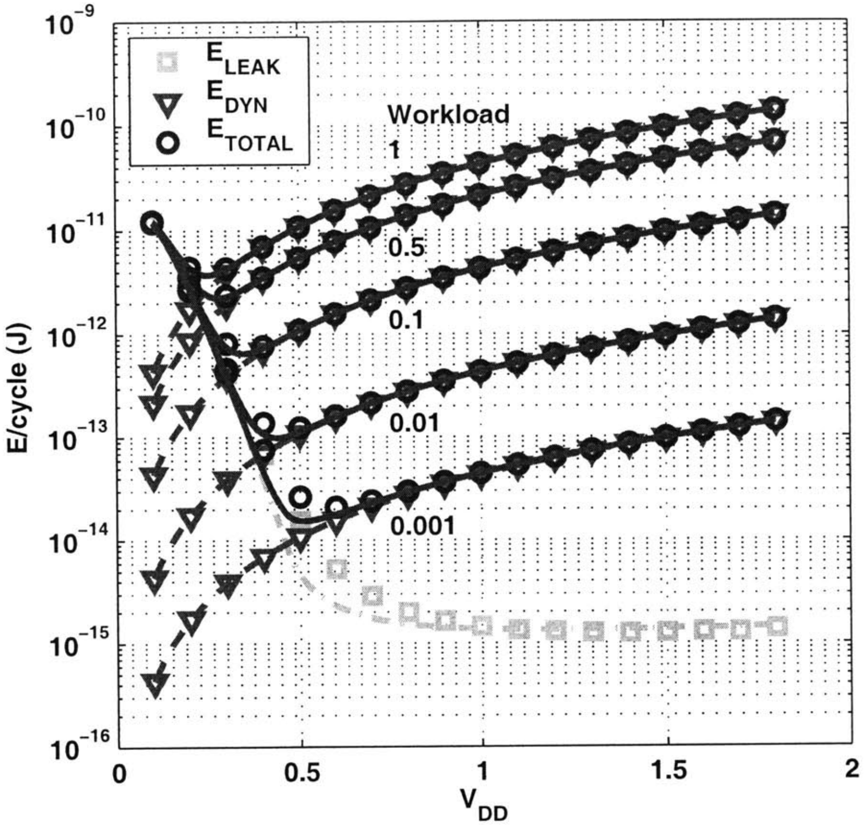


Figure 2-6: Energy versus V_{DD} for changing workload.

magnitude to show the effect. The active energy decreases in proportion to workload because the amount of switched capacitance per operation has decreased. The leakage for the given operation stays constant, however, so the minimum energy point moves to lower energy and higher voltage with decreasing workload.

Duty cycle, d, also can vary widely. A lower duty cycle means that the circuit spends more idle time (e.g. waiting for data but unable to shutdown) for each completed active operation. Consequently, the leakage contribution per operation increases, which corresponds to replacing W_{eff} with W_{eff}/d in equation (2.11). Figure 2-7 shows the impact of a changing duty cycle on the energy characteristics of the FIR filter. Again, a wide range of values is shown for the duty cycle. The longer idle time spent for each operation means that leakage energy increases, but the active

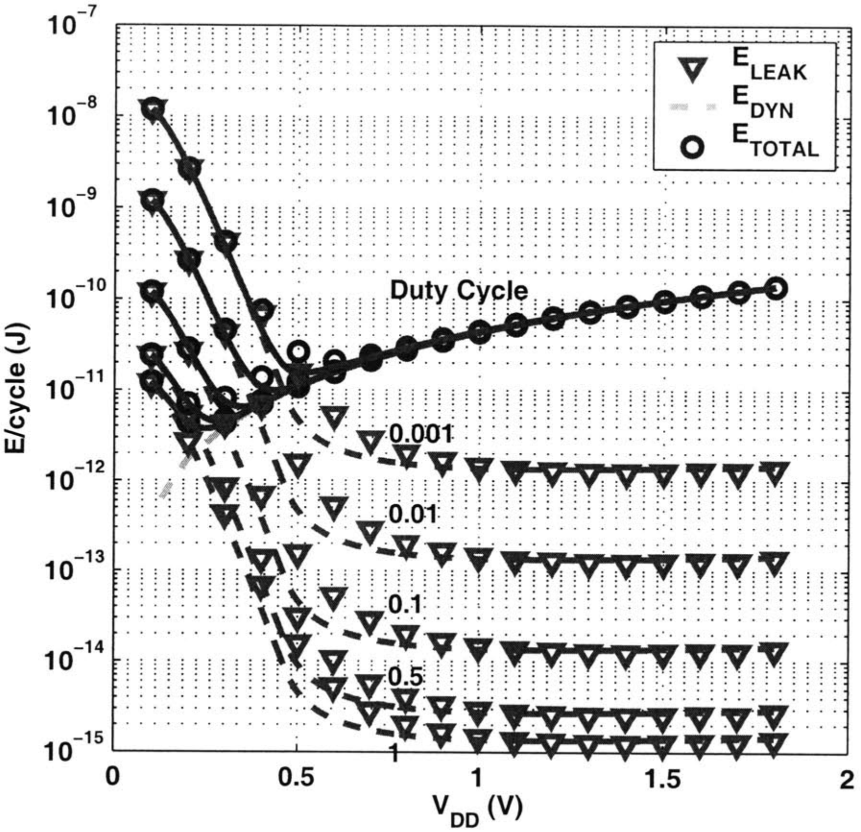


Figure 2-7: Energy versus V_{DD} for changing duty cycle.

energy stays constant. As with workload decreases, the minimum energy point moves to higher voltages, but the total energy increases in this case.

Normalizing to one cycle, we include duty cycle and workload in the analytical model and solve the equation set again to find the optimum V_{DD}, resulting in a new equation for β.

$$\beta = \frac{-2\omega C_{eff}}{\frac{W_{eff}}{d} K C_g L_{DP}} \exp(2) \tag{2.14}$$

Figure 2-8 and Figure 2-9 show the effects of workload and duty cycle on the minimum energy and optimum V_{DD} of the FIR filter. The figures compare the numerical result with the analytical model and with simulation. The supply voltage for the simulations was quantized in 100mV increments. The quantization causes most of the error for values of ω and d close to one. The error in modeled energy at low values of ω and

d occurs because the optimum V_{DD} has exceeded V_T. Thus the assumption of subthreshold operation implicit to the analytical model becomes invalid. The numerical model is also less accurate in that region. The analytical result matches the numerical values quite well until V_{DDopt} nears V_T.

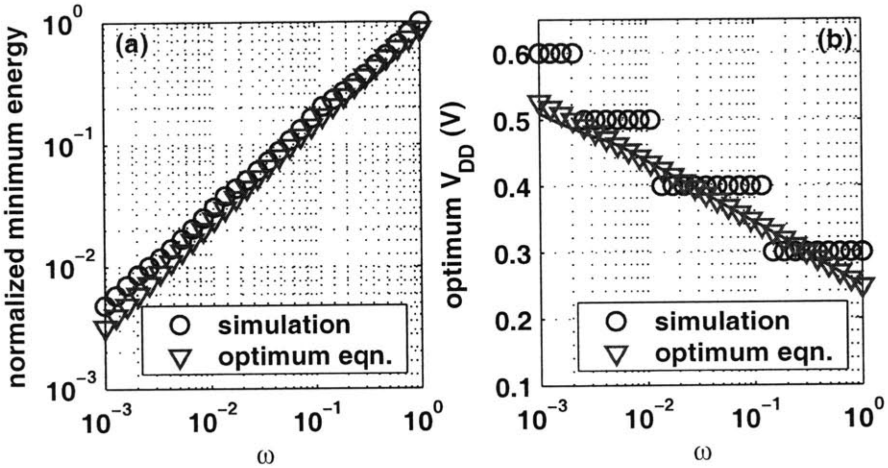


Figure 2-8: Normalized energy (a) and optimum V_{DD} (b) for FIR versus workload, ω. Simulation V_{DD} quantized to 100mV.

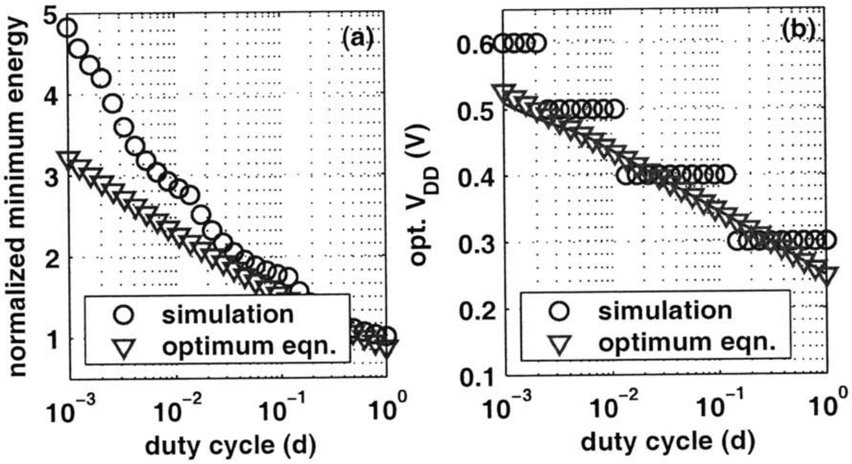


Figure 2-9: Normalized energy (a) and optimum V_{DD} (b) for FIR versus duty cycle, d. Simulation V_{DD} quantized to 100mV.

Large reductions in either ω or d result in increased optimum V_{DD}, but the total energy per operation (normalized to one cycle) decreases as workload decreases and increases when duty cycle decreases. Clearly, knowing the average workload and duty cycle of a circuit can impact the choice of optimum supply voltage. The operational frequency, and thus the data rate, implicitly changes with V_{DD} in these figures. A system in which these parameters vary widely would benefit from closed-loop tracking of the optimum point since Figure 2-8 and Figure 2-9 show a large variation in the minimum energy.

2.3.2 Temperature

The optimum point also depends on temperature. Figure 2-10 shows the effect of temperature on the components of energy. The numerical model shown in Figure 2-10 accounts for temperature dependence by decreasing the effective threshold voltage and decreasing mobility at higher temperatures: $\mu(T) = \mu(T_0)(\frac{T}{T_0})^{-M}$ and $V_T(T) = V_T(T_0) - KT$ [78]. The changes to the numerical model match well with simulation across most of the temperature range, but they slightly underestimate the leakage energy at high temperatures. The lower mobility dominates in strong inversion and leads to slower circuits at high temperatures. In sub-threshold, the lower V_T dominates, and hot circuits grow faster exponentially. The lower V_T that accompanies a temperature increase also raises the leakage current exponentially. This effect appears in the figure at higher V_{DD} where I_{LEAK} dominates E_L. The lower V_T also causes the delay to decrease, countering the increase in I_{LEAK} that is due to lower V_T. Thus, the total effect on E_L and E_T is not so pronounced at lower V_{DD}. Consequently, the total leakage energy does not change quickly near the minimum energy point with V_T. Equation (2.10) shows that V_{DDopt} is linear with temperature, and taking its derivative with respect to temperature gives:

$$\frac{\partial V_{DDopt}}{\partial T} = n\frac{k}{q}\left(2 - \text{lambertW}\left(\frac{-2C_{eff}}{W_{eff}KC_gL_{DP}}\exp(2)\right)\right) \tag{2.15}$$

For the FIR filter, this derivative is only $\sim$ 0.85mV/degree Kelvin, and the plot confirms that the optimum V_{DD} increases by about 75mV over the full temperature

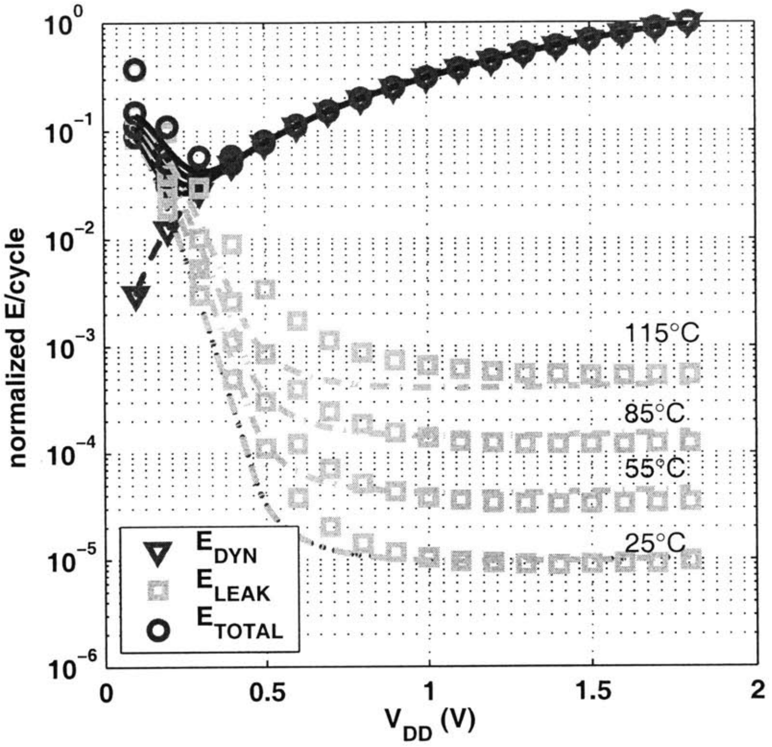


Figure 2-10: Dependence of minimum energy point on temperature shown in simulation (markers) and by the numerical model (lines). Temperature varies from 25°C to 115°C.

range. Refer to Figure 4-8 for measured data from a 90nm test chip showing the impact of temperature on the minimum energy point.

2.3.3 Architecture

The models we have presented allow a quick assessment of the effect of architecture on minimum energy operation. Traditionally, architectural approaches such as parallelism or pipelining can reduce power for a given performance constraint by operating at a lower voltage [53]. For performance constraints that are met at voltages above the optimum, the same conclusion holds for sub-threshold operation. The model can also show how architecture will affect the minimum energy point when performance is not a constraint.

Figure 2-11 shows the numerical model (a) and simulation (b) of a pipelined implementation of the FIR filter. The model does not account for overhead capacitance, leakage, and delay in the pipeline registers, but it shows the general effect of ideal pipelining. As the number of stages increases, L_{DP} decreases and thus reduces leakage energy per operation (E_L). The total energy per cycle thus is reduced, and the optimum V_{DD} moves to the left. The simulation results in Figure 2-11(b) show the same trend, however the overhead active energy makes deep pipelining more costly. The simulation also shows a decrease in active energy for shallow pipelines because of reduced glitching in the multipliers. Thus, shallow pipelining (2-4 stages) can reduce the total energy per operation for a system in a minimum energy scenario.

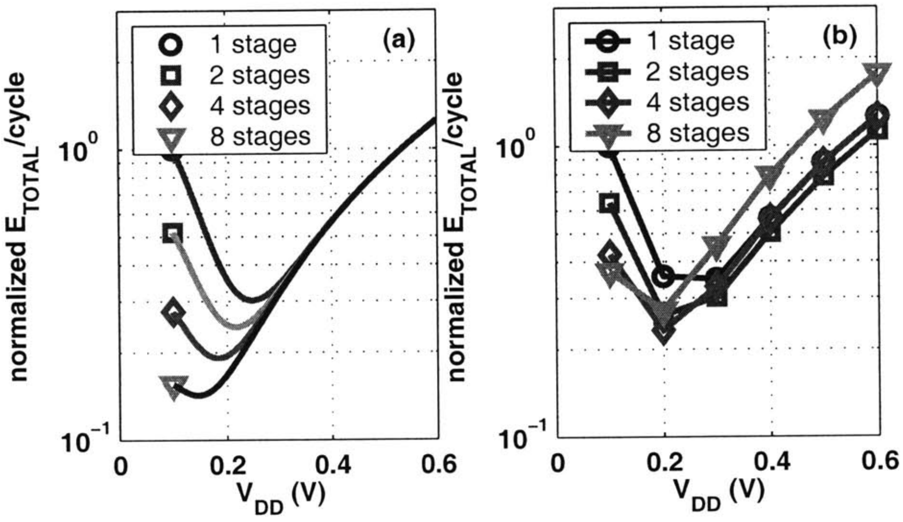


Figure 2-11: Effect of pipelining on minimum energy and optimum V_{DD} for the FIR filter. Ideal pipelines (a) and simulated (b).

In contrast, parallelism cannot reduce the energy per operation, but it can increase the operating frequency at the minimum energy point. Once the optimum V_{DD} is known for a functional unit like the FIR filter, parallel copies of the filter will consume the least energy if they operate at the original minimum energy point. Discounting the overhead of muxing and demuxing indicates that ideal parallelism can increase the operating frequency at the minimum energy point by increasing the number of stages. It cannot, however, decrease the energy per operation because the number of copies of the circuit required to maintain throughput goes up faster than V_{DD} decreases due to the exponential dependence of delay on V_{DD}. Clearly, the overhead of parallelism in a real system will increase active energy and change the minimum energy point.

2.4 Summary and Conclusions

This chapter has examined minimum energy operation for sub-threshold circuits. We have shown that the minimum energy point depends on the technology, the characteristics of the design, and on operating conditions such as temperature, duty cycle, and workload. The optimum V_{DD} for minimizing energy per operation changes over several hundred millivolts when these parameters vary, pointing to the importance of tracking the optimum point or carefully characterizing a design before choosing V_{DDopt}. We introduced an analytical solution for the optimum V_{DD} and V_T to minimize energy for a given frequency in the sub-threshold region. Simulations matched the analytical value and the numerical model within a few percent as long as the sub-threshold assumption was valid.

Chapter 3

Sizing for Minimum Energy

Digital circuits operating in the sub-threshold region provide the minimum energy solution for applications with strict energy constraints. This section examines the effect of sizing on energy for sub-threshold circuits. We show that minimum sized devices are theoretically optimal for reducing energy. A fabricated 0.18μm test chip is used to compare normal sizing and sizing for minimum V_{DD}. Measurements show that existing standard cell libraries offer a good solution for minimizing energy in sub-threshold circuits.

Increasing leakage energy at low supply voltages offsets the reduced active energy and causes a minimum energy point. Many designs exhibit a minimum energy operating point higher than the minimum achievable V_{DD}, and this operating point is a function of several parameters [18][4][75]. In general, designs with larger leakage energy relative to active energy have a higher optimum V_{DD}. This section examines the effect of device sizing on minimum energy operation. After considering theoretically optimal sizing, we explore minimum energy operation for standard cell designs. A fabricated 0.18μm test chip provides measurements for analysis.

3.1 Theoretical Sizing

Sizing influences the energy consumption of a circuit in two primary ways. First, sizing directly affects energy consumption by changing switched capacitance and leakage current. Secondly, sizing affects the minimum voltage at which the circuit functions, which can change the absolute minimum energy point. This section explores these two impacts.

3.1.1 Sizing for Minimum Energy at a Specific V_{DD}

It has been proposed that theoretically optimal minimum energy circuits should use minimum sized devices [56], and first-order equations confirm this result for most cases.

Assuming that the majority of gates in a typical design are sized similarly, a universal increase in transistor sizes will increase both C_{eff} and W_{eff}, raising power. This type of sizing change is unlikely to decrease the critical path delay because the input to output capacitance ratios of gates will stay roughly constant, so the typical assumption of fixed capacitance loads is invalid [79]. Thus, minimum sizing also minimizes energy per operation for most generic circuits. One special case that violates this trend is a circuit with a small number of critical paths relative to the total number of paths. In this case, increased sizes on the critical path can reduce L_{DP} with negligible increases in overall C_{eff} and W_{eff}, lowering E_T.

Minimum sized devices generally minimize energy consumption in sub-threshold for a given V_{DD}. However, sizing also impacts the minimum operating voltage, which can affect the total energy per operation, E_T.

3.1.2 Sizing for Minimum Operating Voltage

Transistor sizing also impacts the functionality of CMOS circuits at low supply voltages. Minimum V_{DD} operation occurs when the pMOS and nMOS devices have the same current (e.g. [57]). Previous efforts have explored well biasing to match the device currents for minimum voltage operation of ring oscillators [62]. Sizing can create the same symmetry in device current. Figure 3-1 shows the minimum voltage for which a ring oscillator maintains 10%-90% voltage swing. The optimum pMOS/nMOS width across all process corners (Typical nMOS, Typical pMOS (TT), Strong nMOS, Strong pMOS (SS), Weak nMOS, Weak pMOS (WW),Strong nMOS, Weak pMOS (SW), and Weak nMOS, Strong pMOS (WS)) is 12. A similar analysis of minimum voltage operation while retaining 10% noise margins gives a lower minimum voltage at the typical corner and a higher worst-case minimum voltage but the same optimum size ratio. Figure 3-2(a) shows the VTCs at the minimum V_{DD} of 70mV for several P/N ratios. The gain is somewhat degraded, but the optimum sized curve is symmetrical and shows good noise margins. Figure 3-2(b) shows the output of a 9-stage ring oscillator at the minimum voltage for the same sizes.

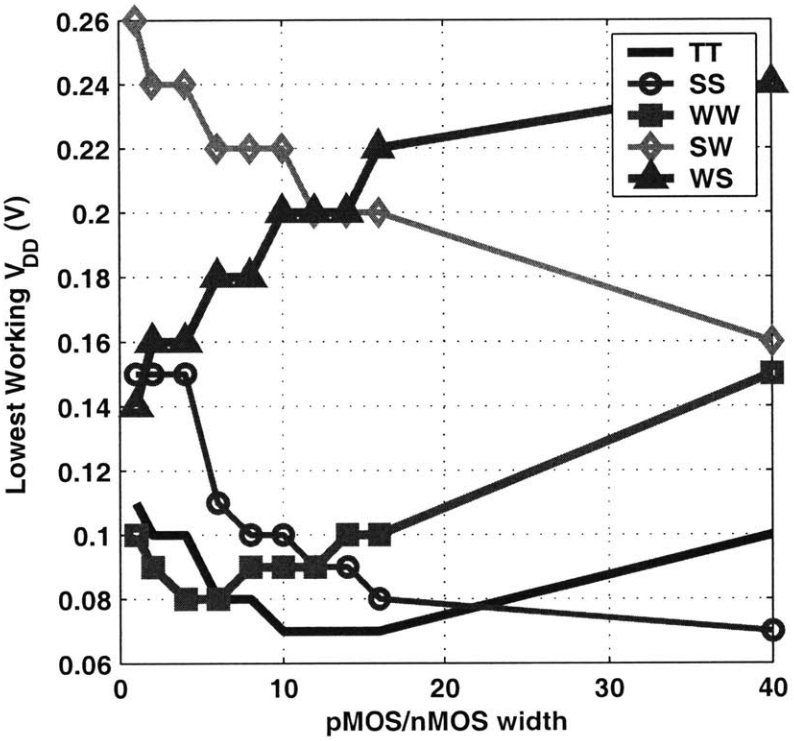
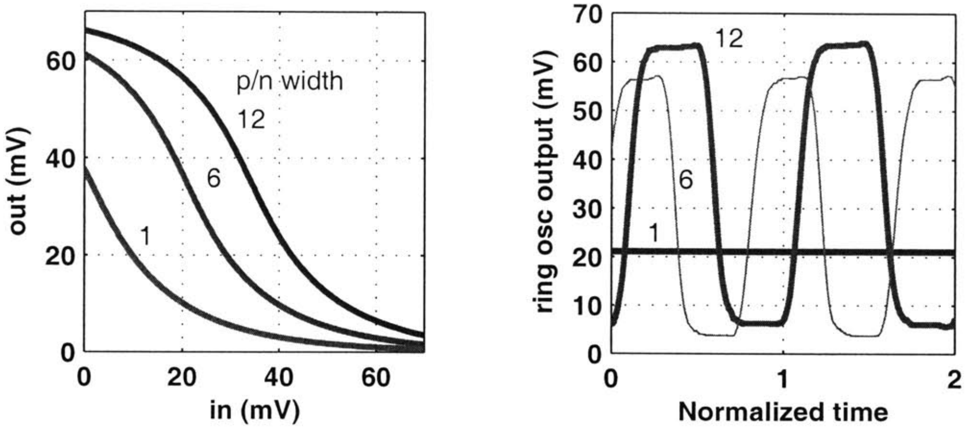


Figure 3-1: Minimum achievable voltage retaining 10%-90% output swing for 0.18μm ring oscillator across process corners (simulation).

Since symmetrical devices give minimum V_{DD} operation, a simple comparison of currents in nMOS and pMOS devices shows the approximate optimum size for minimizing V_{DD}. The switching threshold of a symmetric inverter is $V_M = V_{DD}/2$. Sweeping the width of *on* nMOS and pMOS devices at $V_{DD}/2$ shows the pMOS size



(a) VTCs

(b) Waveforms

Figure 3-2: VTCs (a) and 9-stage ring oscillator output (b) at the minimum V_{DD} for the typical corner (simulation). pMOS/nMOS width ratio of 12 minimizes operational V_{DD} but increases energy consumption.

for which the inverter is balanced. Performing this simulation for V_{DD}=70mV at the typical corner shows that the current in the pMOS matches the nMOS current at a ratio of 12.5. This ratio remains between 11.5 and 12.5 for voltages across the sub-threshold region, matching the result in Figure 3-1.

Sizing according to this ratio allows for operation at lower V_{DD} but increases the energy consumed for a given V_{DD} (equation (2.7)). The energy savings from lowering V_{DD} are at best proportional to V_{DD}^2 if leakage is still negligible. Figure 3-1 shows that the impact of sizing an inverter on the minimum supply voltage is only 60mV, producing best-case energy savings of $0.20^2/0.26^2 = 0.6$X due to voltage reduction. This improvement is not worthwhile if all pMOS devices are increased in size by 12X. The exponential dependence of current in sub-threshold on V_T means that sizing is a less effective knob for 'fixing' circuits. This explains why the knob has to turn so far to match the transistor current that is imbalanced by a small difference in V_T. Thus, minimum sized devices are theoretically optimal for reducing energy per operation when accounting for the impact of sizing on voltage and energy consumed [75]. Process variation in deep submicron processes imposes one restriction to applying this rule blanketly. The sigma for V_T variation due to random doping fluctuations is proportional to $(WL)^{-\frac{1}{2}}$, so minimum sized devices produce the worst case random V_T mismatch. Statistical analysis is necessary to confirm functionality in the face of process variation, and some devices might need to increase in size to ensure acceptable yield.

3.1.3 Ratioed Circuits

Generally, static CMOS circuits are robust for sub-threshold operation. The large sensitivity of sub-threshold circuits to process variation makes some circuit designs more suitable than others. Ratioed circuits can create problems with functionality in sub-threshold.

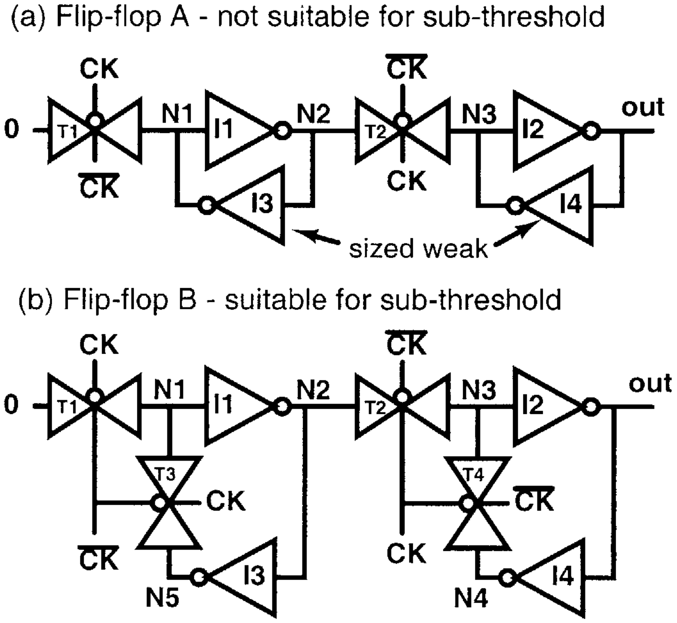


Figure 3-3: Two flip-flops for evaluation for sub-threshold operation. Flip-flop A has ratioed feedback, and Flip-flop B does not. Ratioed circuits cannot function across process corners in sub-threshold at the minimum energy voltage because the exponential dependence of current on V_T becomes more important than sizing. Cutting the feedback loop for writing a latch is robust across all corners for operation at the minimum energy voltage.

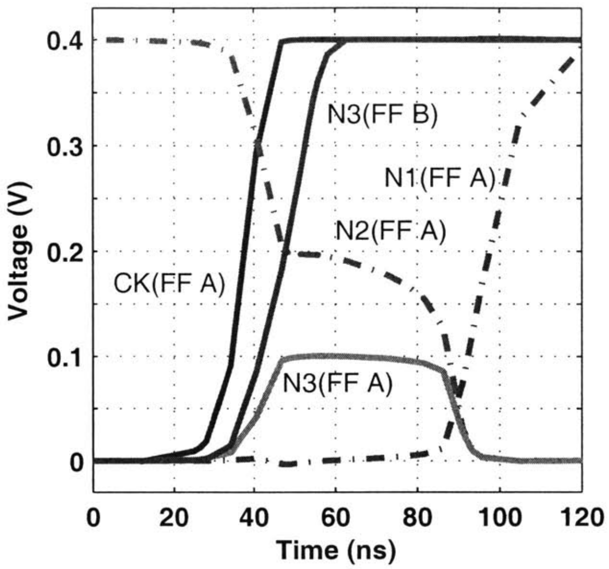


Figure 3-4: Plot showing the failure of ratioed Flip-flop A at the strong nMOS, weak pMOS corner at 400mV. The master stage fails to write the slave stage and is itself overwritten.

For example, Figure 3-3 demonstrates an important consideration about circuit selection for sub-threshold. The figure shows a ratioed flip-flop (Flip-flop A) and a flip-flop with transmission gates for cutting off feedback in the latches (Flip-flop B). The ratioed flip-flop uses devices sized to be strong (wide) on the critical path and inverters sized to be weak for the feedback in the latches. This approach depends on the sizing ratios in the circuit to guarantee functionality, and it works well above-threshold. In sub-threshold, however, current depends exponentially on V_T. This makes process variation and local device variation quite significant relative to device sizes. Figure 3-4 shows a plot of Flip-flop A failing to write node N3 to a one at the strong nMOS, weak pMOS corner. The pull-up path through I1 and T2 is weak because of the higher pMOS threshold voltage and cannot overcome the pull-down path through I4 and T2 that is made strong by the lower nMOS threshold voltage. Instead, the pull-down path overcomes I1 and actually flips the state of the master stage by pulling N2 to zero. Thus, Flip-flop A fails below 450mV at this corner, preventing the circuit from reaching its minimum energy voltage. Massively up-sizing the pMOS devices can correct this problem, but a better choice is to eliminate the ratioed fight by adopting Flip-flop B. This flip-flop cuts off the feedback path before writing a latch, allowing the V_{GS} applied to on-transistors to increase current beyond any device off-currents even at process corners.

In general, the strong dependence of sub-threshold current on V_T and temperature makes the ratio of sizing inadequate for compensating across the full process corner and temperature ranges in sub-threshold operation. As a result, non-ratioed circuit styles provide more robust functionality in sub-threshold.

3.2 Standard Cells

Standard cell libraries aid digital circuit designers to reduce the design time for complex circuits through synthesis. Most standard cell libraries focus on high performance, although including low power cells is becoming more popular [79]. Lower power cells generally use smaller sizes. One standard cell library geared specifically for low power uses branch-based static logic to reduce parasitic capacitances and a reduced set of standard cells. Eliminating complicated cells with large stacks of devices and using a smaller total number of logic functions was shown to reduce power and improve performance [80]. Standard cell libraries have not been designed specifically for sub-threshold operation. This section evaluates the performance of a 0.18μm standard cell library in sub-threshold operation.

We use an 8-bit, 8-tap, parallel, programmable FIR filter as a benchmark to compare normal standard cells with cells that are modified to allow operation at the minimum energy point across all process corners. Figure 3-5 shows the minimum operating voltages for the different standard cells appearing in a normal synthesis of the FIR filter. The TT and worst-case (Fast nMOS, Slow pMOS (FS) and Slow nMOS, Fast pMOS (SF)) process corners are shown. All of the cells operate at 200mV at the typical corner, showing the robustness of static CMOS logic. Additionally, most of the cells operate at 300mV in the worst case. The cells which exhibit the worst case (failing below 400mV) are flip-flops and complex logic gates with stacks of

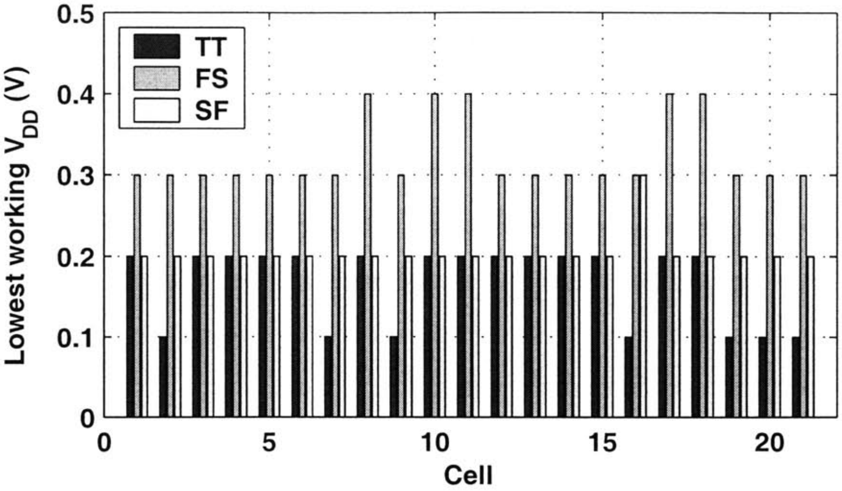


Figure 3-5: Standard cell functionality in synthesized FIR filter using normal cell selection over process corners (simulation).

series devices (e.g. And/Or/Invert (AOI)). We eliminated the problematic cells by preventing the synthesis tool from selecting logic gates with large device stacks and by re-sizing a few offending cells such as the flip-flop and full adder.

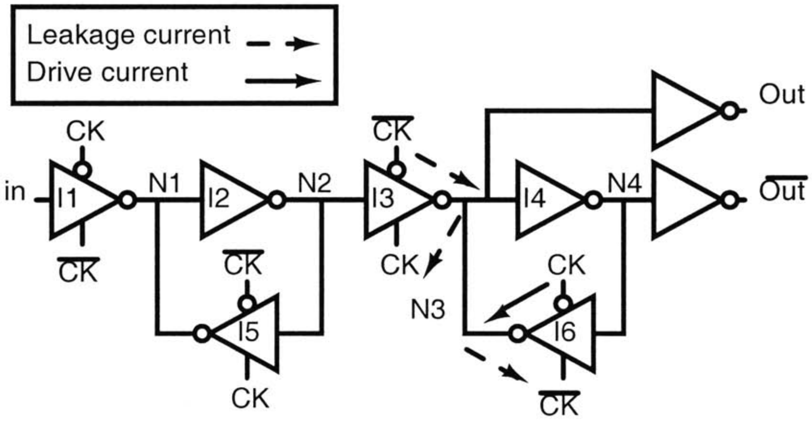


Figure 3-6: Standard cell flip-flop at worst-case failure point where $CK = 0$ at FS corner (fast nMOS, slow pMOS).

Figure 3-6 shows a schematic of the D-flip-flop. In the standard implementation, all of the inverters use small nMOS and only slightly larger pMOS devices except I3, which is several times larger to reduce Clock-to-Q delay. At the FS corner, the narrow

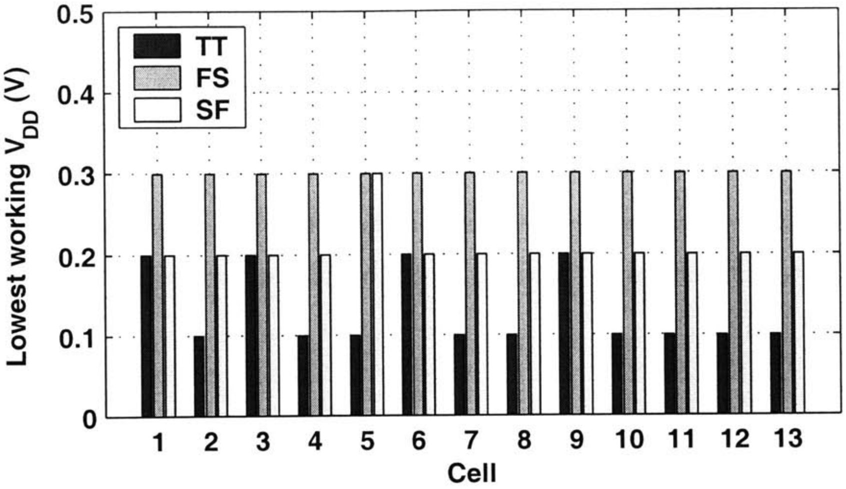


Figure 3-7: Standard cell functionality in synthesized FIR filter using cells sized to minimize V_{DD} over process corners (simulation)

pMOS in I6 cannot hold N3 at a one when CK is low. This is because the combined, strong *off* current in the nMOS devices in I6 and I3 (larger sized) overcomes the weakened, narrow pMOS device in I6. Tying back to the ring oscillator in Figure 3-1, the combined nMOS devices create an effective pMOS/nMOS ratio that is less than one. To prevent this, we reduced the size of I3 and strengthened I6. Clearly, the larger feedback inverter creates some energy overhead. However, the resized flip-flop can operate at 300mV at all process corners in simulation. Figure 3-7 shows the lowest operating voltage for the cells in the minimum-V_{DD} FIR filter. The number of cell types has reduced, and all of the cells work at 300mV across all corners. The next section uses test chip measurements to compare the filter sized for minimum V_{DD} with the normal filter.

3.3 Measured Results from Test Chip

A 0.18μm, 6M layer, 1.8V, 7mm^2 test chip was fabricated to measure the impact of sizing on minimum energy operation of standard cells. The test chip features two programmable 8-bit, 8-tap FIR filters. Both filters produce non-truncated 19-

bit outputs. The first filter was synthesized using the unmodified synthesis flow and normal cells (Figure 3-5). The second filter was synthesized using the modified flow in which some cells were omitted and some cells were resized to minimize V_{DD} (Figure 3-7). Both filters can operate using an external clock or an on-chip clock generated by a ring oscillator that matches the respective critical path delay of the filters. Input data comes from an off chip source or from an on-chip Linear Feedback Shift Register (LFSR).

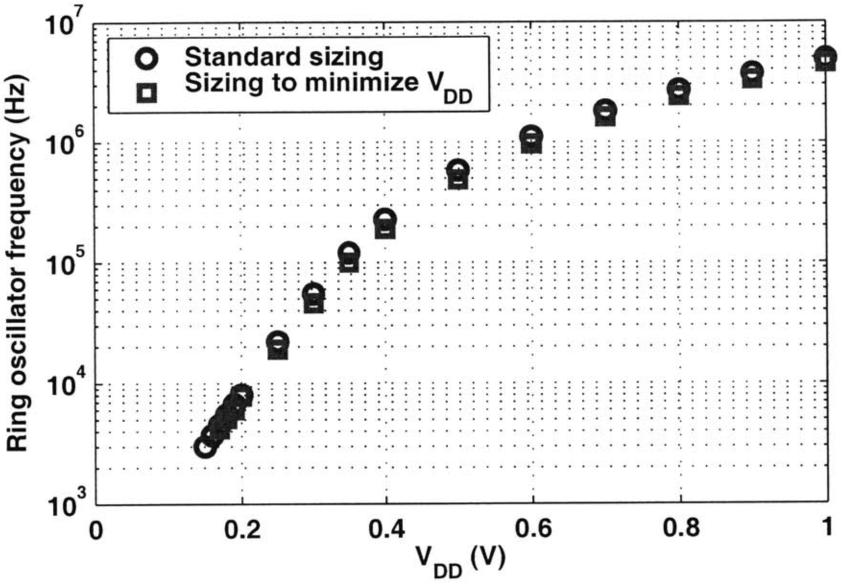


Figure 3-8: Measured performance of programmable FIR filters on the test chip. Standard FIR is 10% faster than the minimum-voltage FIR.

Figure 3-8 shows the measured performance versus V_{DD} for the two filters using their respective critical path ring oscillators and the LFSR data to produce one pseudorandom input per cycle. The minimum-V_{DD} filter exhibits a 10% delay penalty over the standard filter. Both filters operate in the range of 3kHz to 5MHz over V_{DD} values of 150mV to 1V. Both filters are fully functional to below 200mV.

Figure 3-9 shows an oscilloscope plot of the standard filter working correctly at $V_{DD} = 150$mV. The clock in this plot is produced by the ring oscillator on-chip. The reduced drive current and large capacitance in the output pads of the chip cause the

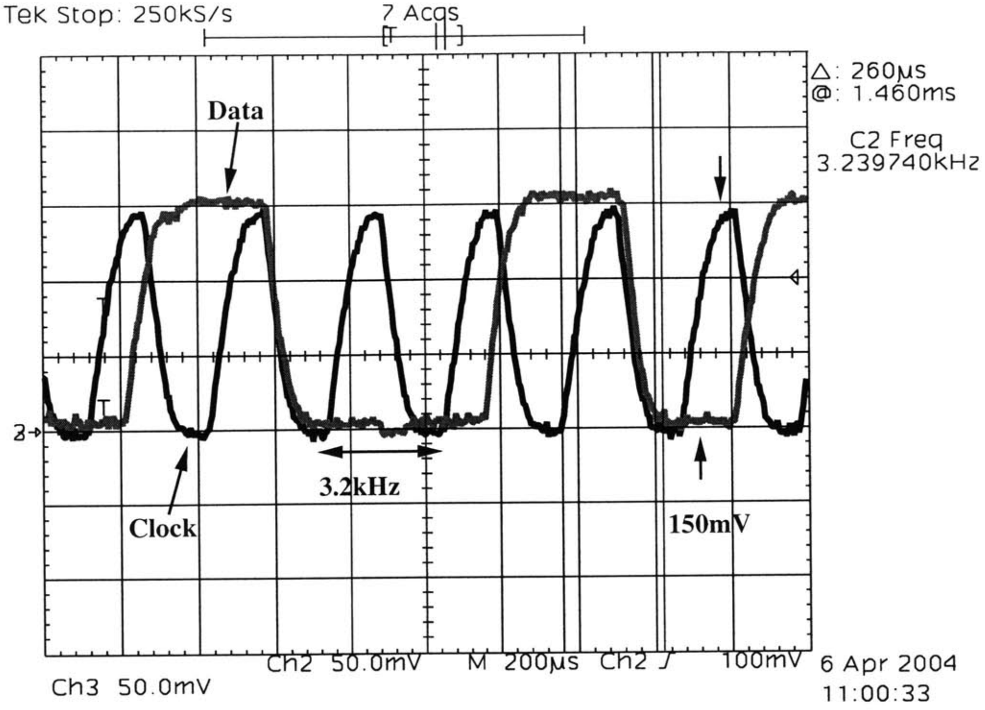


Figure 3-9: Oscilloscope plot from the test chip showing V_{DD} = 150mV filtering operation with ring oscillator clock at 3.2kHz.

slow rise and fall times in the clock, but the signal is still full swing. One bit of the output is shown.

Figure 3-10 shows the measured total energy per output sample of the two FIR filters versus V_{DD}. The solid line is an extrapolation of $C_{eff}V_{DD}^2$ for each filter, and the dashed lines show the measured leakage energy per cycle. Clearly, both filters exhibit an optimum supply voltage for minimizing the total energy per cycle. Within the granularity of the measurements, the optimum V_{DD} is 250mV for the standard FIR, which matches the analytical solution derived in Section 2.2.1. The optimum V_{DD} is 300mV for the minimum-V_{DD} FIR. There is a measured overhead energy per cycle of 50% in the filter sized for minimum V_{DD}. The figure also shows the simulated worst-case minimum V_{DD} for the two filters (cf. Figure 3-5, Figure 3-7). Accounting for overhead at the worst-case minimum V_{DD}, the minimum-V_{DD} FIR offers a reduction in total energy of less than 10% at the worst-case process corner,

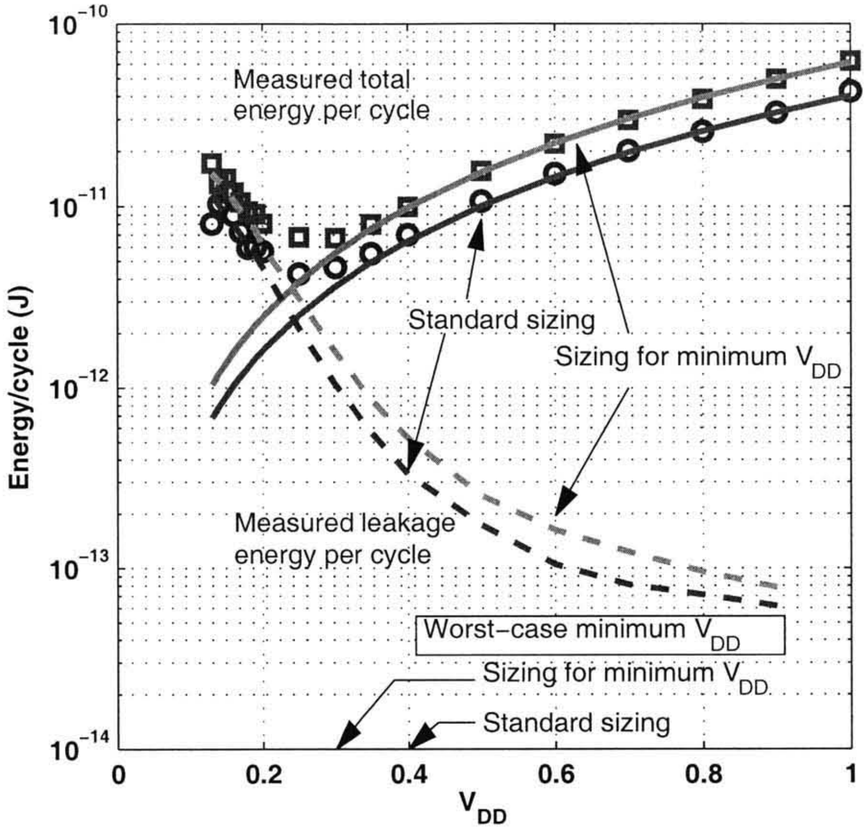


Figure 3-10: Measured energy per operation of the FIR filters on the test chip.

but this improvement comes at a cost of 50% at the typical corner.

Simulations show that the measured overhead cost in the minimum-V_{DD} filter primarily results from restricting the cell set that the synthesis tool could use. Since the tool was not optimized for the smaller set of cells, we did not see the improvements that are possible through this approach [80]. Using only sizing to create the minimum V_{DD} filter would have decreased the overhead. However, the shallow nature of the optimum point in Figure 3-10 shows that the unmodified standard cell library does not use much extra energy by failing at a higher V_{DD} at the worst-case corner. Thus, existing libraries provide good solutions for sub-threshold operation. Simulation shows that a minimum-sized implementation of the FIR filter has 2X less switched capacitance than the standard FIR, so a mostly minimum-sized library theoretically would provide minimum energy circuits.

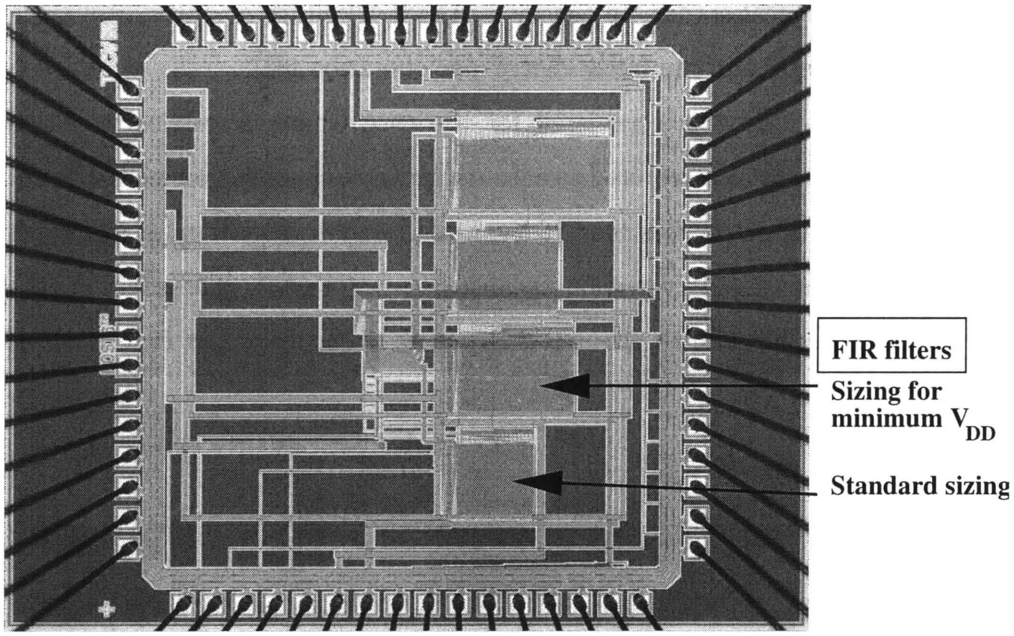


Figure 3-11: Annotated die photo of 0.18μm sub-threshold FIR test chip.

3.4 Summary and Conclusions

This chapter has examined device sizing for sub-threshold operation. For typical circuits and modern technologies, the optimum supply voltage for minimizing power is higher than the failure point for minimum sized devices at the typical corner. Thus, minimum sized devices are theoretically optimal for minimizing power. Even if the minimum energy point for a certain process corner or unusual circuit occurs at a supply voltage where minimum sized devices cannot function, the shallow nature of the optimum prevents up-sizing to reduce the minimum possible operating voltage from being worthwhile. Again, the primary reason that sizing is less effective in sub-threshold is that it impacts current linearly. It is thus less effective when compensating for V_T differences that have an exponential impact on current. Measurements from a test chip, shown in Figure 3-11, confirm that existing static CMOS standard cell libraries function well in sub-threshold. Resizing or restricting cell usage in such libraries can lower the worst-case minimum V_{DD}, but the overhead increases energy consumption at the typical corner. In theory, a standard cell library primarily using minimum-sized devices would minimize energy per operation [81].

The work in this section is completed for the case where process variation is ignored. Remaining analysis is required to account for the impact of process variation. Specifically, minimum sized devices provide the worst possible deviation for mismatch in V_T caused by random doping variation because $\sigma V_T \propto (WL)^{-\frac{1}{2}}$. The effect of such variation will place a constraint on sizing in some cases to ensure proper functionality.

Chapter 4

Ultra-Dynamic Voltage Scaling

Dynamic Voltage Scaling (DVS) has become a standard approach for reducing power when performance requirements vary. DVS systems lower the frequency and voltage together to reduce power when lower performance is allowed. DVS was introduced as a method to lower power when performance requirements change [82]. In the first true DVS implementation, a critical path replica is used in a feedback loop to adjust the supply voltage to the lowest value that allows the delay to match a given reference frequency [82]. DVS now appears in commercial processors such as, for example, the Intel XScale [83], IBM PowerPC [84], and the Transmeta Crusoe processor [85].

Voltage dithering was proposed as a low overhead implementation of DVS to provide near-optimum power savings using only a few discrete voltage and frequency pairs [5]. The savings are only achievable if the voltage and frequency can change on the same time scale as the altering workload. Previous implementations apply voltage dithering to entire chips and require many microseconds to change operating voltage [5][86]. This chapter describes a 90nm test chip that demonstrates the proposed concept of Local Voltage Dithering (LVD) and couples LVD with sub-threshold operation to achieve Ultra-Dynamic Voltage Scaling (UDVS) [87]. We provide measurements of the effect of temperature on minimum energy operation for the 90nm test chip.

Our work extends the state-of-the-art in several ways. First, LVD improves on existing voltage dithering systems by taking advantage of faster changes in workload and by allowing each block to optimize based on its own workload. Additionally, we show that the time and energy overhead of LVD are small based on measurements from the test chip. Secondly, UDVS provides a practical method for extending DVS into the sub-threshold region. For many emerging energy-constrained applications, lowering energy consumption is the primary concern under most conditions. Thus, operating at the minimum energy point conserves energy at the cost of lower performance (frequency). This type of application works at the minimum energy point primarily, and only jumps to higher performance voltages in rare cases. We have shown the effectiveness of UDVS for this scenario.

4.1 DVS and Local Voltage Dithering

Many signal processing systems process blocks of data that arrive at some regular rate, and sometimes the amount of data to process is less than the maximum amount. This corresponds to a fixed-throughput system whose workload requirements change on a block-to-block basis in a time varying fashion. Examples of this type of application include MPEG video processing and FIR filtering with a variable number of taps [5]. In the video processing example, the maximum workload corresponds to a scene change in the video sequence. In this case, the entire new frame of data requires processing since it is completely different from the previous frames. In the absence of scene changes, new frames of data may not differ significantly from the previous frame, so only a small section of the new frame requires processing. This case represents a reduced workload for the system. The workload of the system measures the amount of processing required for a given block of data, and the rate is simply the normalized processing frequency [5]. In a system without buffering, the lowest allowable rate equals the workload. If buffering is possible, then there are different strategies involving operation at different rates that can correctly perform the required processing on a block with a given workload by ensuring that the average rate equals the workload. There are many applications where workload varies with time [88], and policies for setting the rate based on incoming data have been explored [89][90][91].

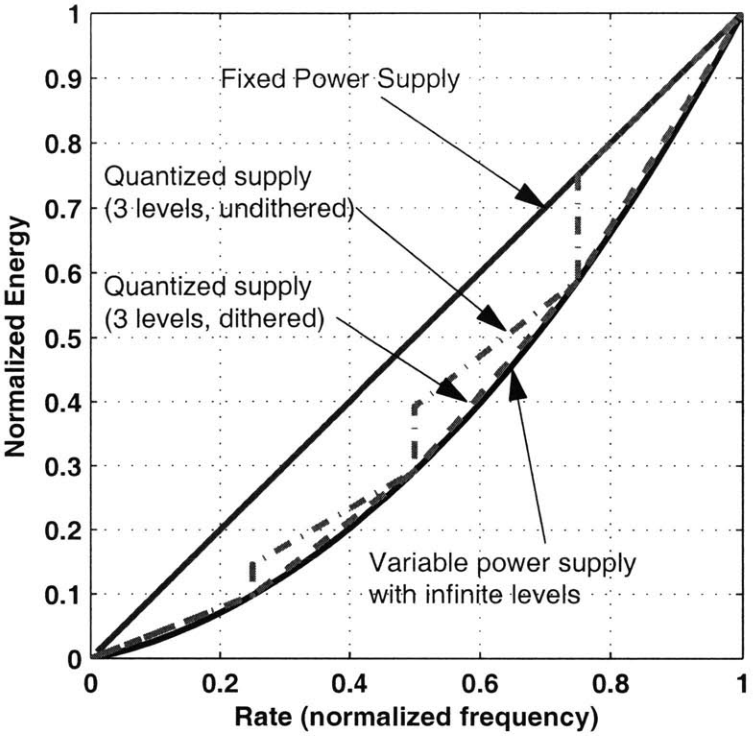


Figure 4-1: Theoretical energy consumption versus rate for different power supply strategies [5].

Figure 4-1 shows four approaches to power supply management for reducing energy consumption when the workload varies [5]. It plots the required rate of the system versus the normalized energy required to process one generic block of data. The most straightforward method for saving energy when the workload decreases is to operate at the maximum rate until all of the required processing is complete and then to shutdown. This approach only requires a single power supply voltage (corresponding to full rate operation), and it results in linear energy savings. The fixed power supply curve in Figure 4-1 assumes ideal shutdown (i.e. - no shutdown power). A variable supply voltage with infinite allowable levels provides the optimum curve for reducing energy. This curve in Figure 4-1 corresponds to theoretically ideal DVS according to the model in [5] where velocity saturation is omitted. When velocity saturation occurs, the energy savings for ideal DVS increase because the performance does not decrease as quickly for the same change in V_{DD} [86].

4.1.1 Voltage Dithering

One method that avoids the problem of creating an infinite number of supply voltages is to use quantized supply voltages. In Figure 4-1, three levels of supply voltage quantization are used with two different policies. The undithered policy simply selects the lowest supply voltage for which the rate exceeds the desired rate, operates at that rate and voltage until all of the data in the block is processed, and then shuts down. This results in the stair-step energy characteristic. A better method is called voltage dithering [5]. The basic idea behind voltage dithering is to divide the computation of one block of data between operation at the quantized supply voltage and rate pairs that occur above and below the desired average rate. The energy profile for dithering between quantized voltage supplies linearly connects the quantized rate and energy pairs on the plot. Assuming that the desired rate of operation for a block, R_{BLOCK}, lies between two quantized rates, R_{LOW} and R_{HIGH}, then:

$$E_{BLOCK} = E_{LOW}\left(\frac{R_{HIGH} - R_{BLOCK}}{R_{HIGH} - R_{LOW}}\right) + E_{HIGH}\left(\frac{R_{BLOCK} - R_{LOW}}{R_{HIGH} - R_{LOW}}\right) \qquad (4.1)$$

where E_{HIGH} and E_{LOW} are the normalized energies consumed for processing a block at R_{HIGH} and R_{LOW}, respectively.

Figure 4-2 shows an example comparing voltage dithering with fixed and variable supply approaches. This example uses two quantized voltages that provide full rate and half rate operation. For a desired rate of 0.6, the fixed supply approach operates at the highest rate and full power for 0.6 of the full block time (Figure 4-2(b)). Ideal variable voltage operation provides exactly 0.6 rate at the best possible energy (Figure 4-2(d)) Voltage dithering gives an average rate of 0.6 by operating for 20% of the block time at full rate and for 80% of the time at 0.5 rate. The resulting energy consumption is thus averaged between the two quantized points and falls on the connecting line (Figure 4-2(c)). This approach allows a good approximation of the optimum energy profile with less overhead.

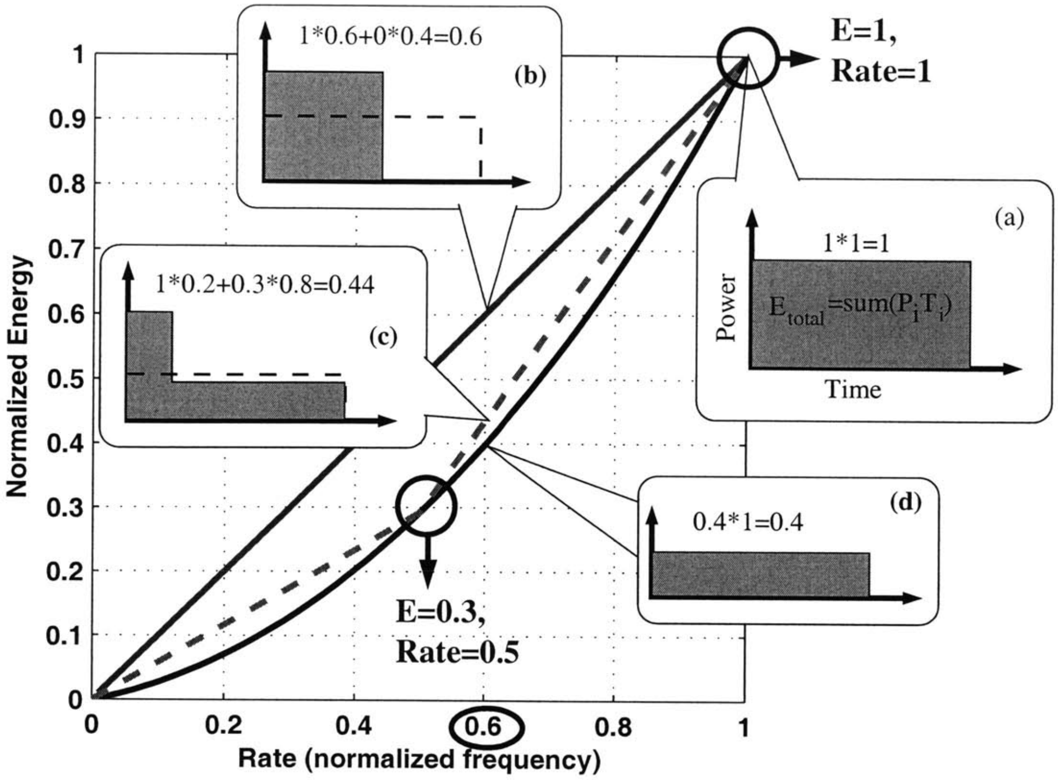


Figure 4-2: Voltage dithering example for 0.6 rate normalized to full rate (a). Example shows fixed supply (b), voltage dithering (c), and ideal variable supply (d).

Implementations of systems that use voltage dithering apply it monolithically to an entire chip. The system in [5] uses an on-chip variable dc-dc converter to dither the voltage supplied to the entire chip. A chip containing header switches was used in [86] to select the voltage supplied to a different chip (off-the-shelf processor). A similar off-chip voltage hopping approach is used in [92] for a zero V_T processor in fully depleted Silicon on Insulator (SOI). These implementations have shown the effectiveness of voltage dithering to save energy for high performance applications with variable workload.

4.1.2 Local Voltage Dithering

Applying voltage dithering at the local level provides several key advantages over previous chip-wide applications. We have proposed local voltage dithering (LVD) to improve upon chip-wide voltage dithering. This section discusses the advantages of LVD and describes a test-chip that demonstrates these improvements.

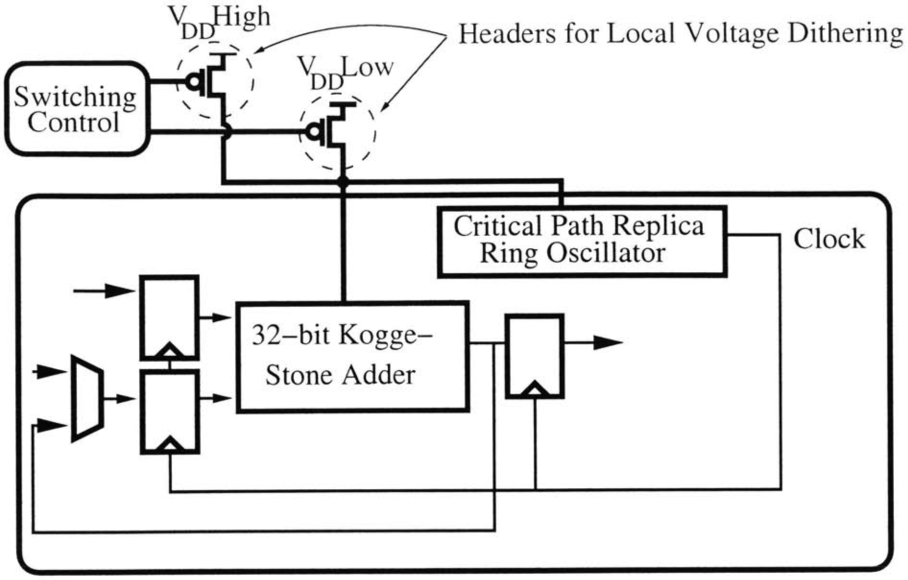
Previous chip-wide implementations using voltage dithering report that the transition between two different supply voltages takes hundreds of micro-seconds [5][86]. This prevents the system from achieving any energy savings for faster changes in the workload. Dividing up the power supply grid into local regions reduces the capacitance that must be switched when the voltage supplied to a local block needs to change. This allows for faster changes in supply voltage with lower transitional energy and permits energy savings for changes in workload on the same timescale.

Chip-wide voltage dithering also restricts the extent to which varying workload may be leveraged because it must account for the highest workload from all of the blocks across the entire chip. For example, suppose a simple chip contains two large blocks. If one block has a workload of 0.9 and the other block has a workload of 0.2, then chip-wide voltage dithering must ensure that the block with the higher workload completes its work. Since both blocks share the dithered voltage supply, they both are forced to operate at the average rate of 0.9. Even if the less active block shuts down (e.g. clock gates) after completing its processing, it still uses more energy than if it could voltage dither based on its own workload. The energy savings that are lost by using chip-wide voltage dithering only increase with more blocks and wider differences between the maximum and minimum workloads. In contrast, LVD lets each block operate according to its own workload.

Our implementation of LVD uses embedded power switches (pMOS header devices) to toggle among a small number of voltage levels at the local block level. One advantage of this implementation approach is that the local dithering switches can be turned off to provide fine-grained power gating essentially for free.

4.2 UDVS Test Chip

We have implemented a test chip in 90nm bulk CMOS to demonstrate LVD and UDVS. This section describes the test chip architecture and provides measured results



V_{DD}High
Headers for Local Voltage Dithering
V_{DD}Low
Switching Control
Critical Path Replica Ring Oscillator
Clock
32–bit Kogge–Stone Adder

Figure 4-3: Block diagram of voltage dithered adder and critical path replica using two local header switches for local voltage dithering (LVD).

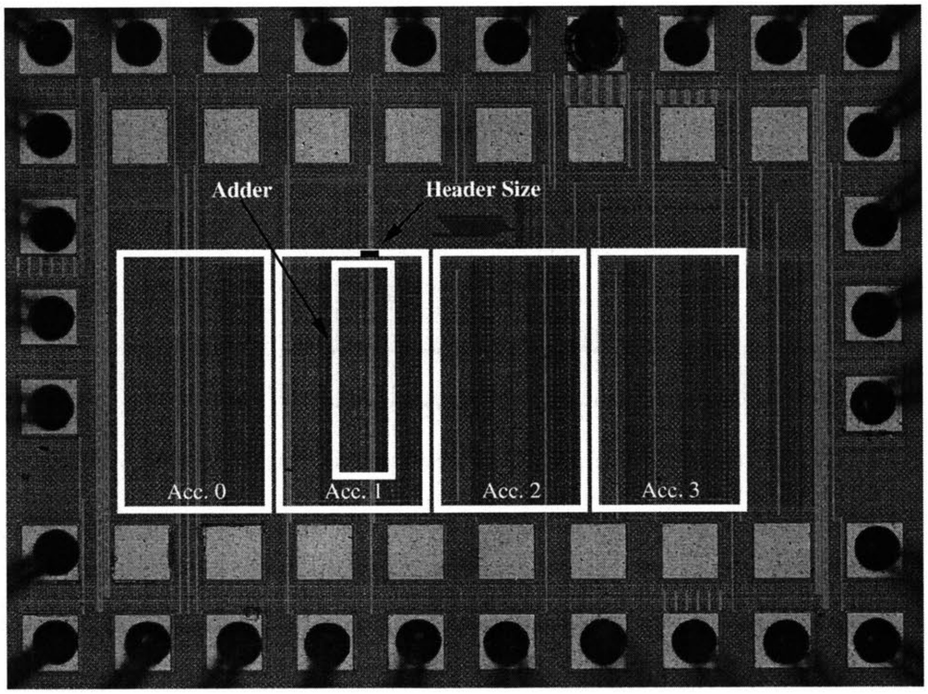


Figure 4-4: Annotated die photograph showing accumulators with 0, 1, 2 and 3 headers. The size of one header is highlighted for reference.

4.2.1 UDVS Test Chip Architecture

Figure 4-3 shows the primary block used for testing LVD on the test chip. The circuit of interest is a 32-bit Kogge-Stone adder that can be configured as an accumulator for testing. In this figure, two pMOS header switches select between a high supply voltage (V_{DDH}) and a low supply voltage (V_{DDL}) for the adder block. Other adders on the chip have different numbers of header devices. A critical path replica ring oscillator shares the same dithered voltage supply as the adder and sets the frequency of the clock based on the selected voltage. The die photo in Figure 4-4 shows the accumulators with different numbers of header switches used for testing, and the approximate area of a single header switch is highlighted for reference.

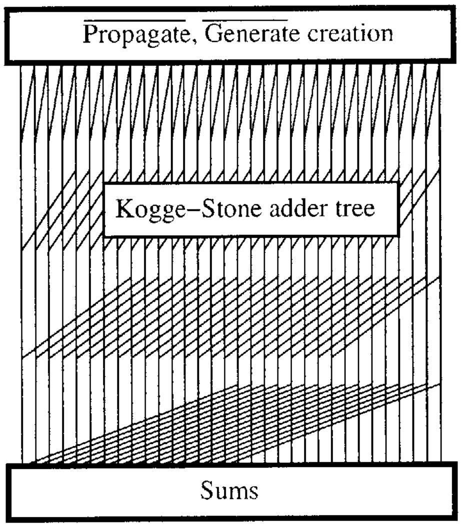

$\overline{\text{Propagate}}$, $\overline{\text{Generate}}$ creation
Kogge–Stone adder tree
Sums

Figure 4-5: Schematic of adder circuits. Kogge-Stone-based tree with inverting stages of dot operators (at each reconvergence of the tree).

Placing a pMOS header switch in series with the power supply increases the delay of the circuit because of the voltage drop across the on resistance of the header. This effect is well-known and thoroughly analyzed in the context of power gating approaches such as multi-threshold CMOS (MTCMOS). Numerous methods for sizing such header devices are available, and most of them are designed to ensure that the circuit never exceeds some delay penalty relative to the circuit without any headers.

The header switches on the test chip are sized to keep the delay penalty less than 10%.

Figure 4-5 shows the architecture, and Figure 4-6 shows the circuit schematics for the adder block on the test chip. The inverse of the propagate and generate signals are calculated in the first stage, and these results are applied to the adder tree. Each reconverging point in the tree has a "dot operator" circuit that calculates the propagate and generate values for that stage. Each stage in our implementation is inverting, so the two flavors of dot operator are shown in the critical path schematic.

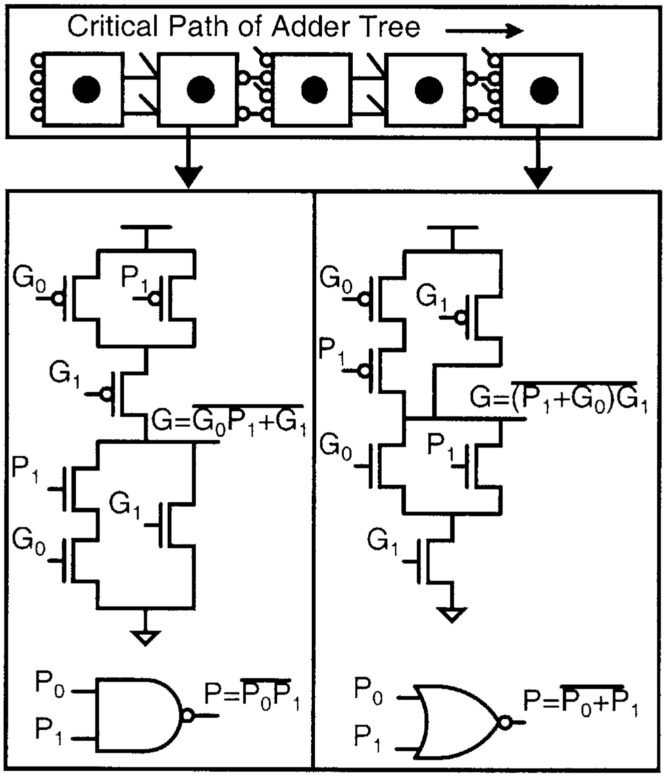


Figure 4-6: Circuits for Kogge-Stone adder. Inverting stages of dot operators are in series along the critical path. Circuits do not require large stacks of transistors, which degrade sub-threshold operation.

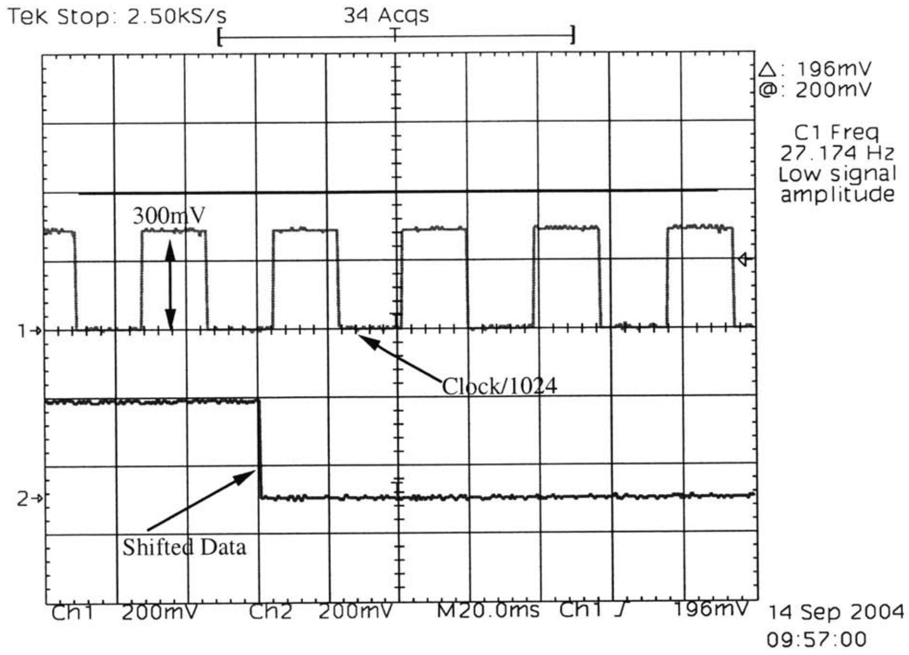


Figure 4-7: Oscilloscope plot showing the clock and data from the 90nm test chip operating at 300mV, just below the minimum energy point. The adders functioned to 200mV.

4.2.2 Measurements

Since UDVS scales the supply voltage from the full V_{DD} down to the optimum V_{DD} for minimum energy, a UDVS system must consist of circuits that can function in the sub-threshold region. This test chip uses static CMOS circuits to ensure robust sub-threshold operation. The adder blocks on the test chip operate to below 200mV. Figure 4-7 shows an oscilloscope plot of the adder on the 90nm test chip operating in sub-threshold at 300mV, just below the minimum energy voltage.

The minimum energy per operation point measured for the adder appears in Figure 4-8 at $V_{DD} = 330\text{mV}$ ($f = 50\text{kHz}$) and 0.1pJ per addition for 25^oC. Figure 4-8 also shows the measured effect of temperature on the total energy per cycle and leakage energy per cycle. An increase in temperature lowers the mobility of MOSFETs and decreases the threshold voltage according to: $\mu(T) = \mu(T_0)(\frac{T}{T_0})^{-M}$ and $V_T(T) = V_T(T_0) - KT$ [78]. For above-threshold operation, the decreased mobility dominates, and circuits slow down as they heat up. The leakage energy increases quickly with temperature for $V_{DD} > V_T$ because of the exponential dependence on the lower threshold voltage. In the sub-threshold region, however, the increased current also decreases the cycle delay, which causes the higher leakage currents to integrate over a shorter cycle time. As a result, the leakage energy does not change enough with temperature to greatly impact the optimum supply voltage. Figure 4-8 shows that the measured effect of temperature on the minimum energy point is small, validating the model in [75] and the analysis in Section 2.3.2. Figure 4-9 shows the measured frequency of one of the critical path ring oscillators on the test chip versus V_{DD} and temperature, confirming the increase of performance at higher temperatures in the

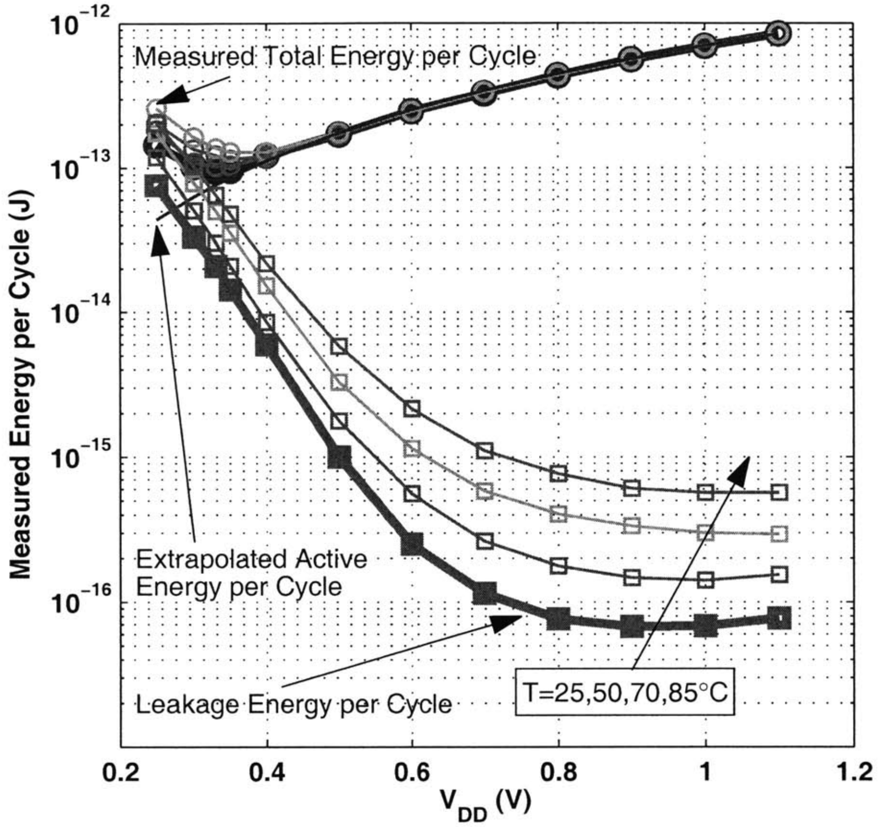


Figure 4-8: Measured energy per cycle in the sub-threshold region for input activity of one. Minimum energy point occurs at 330mV (50kHz) and 0.1pJ per operation at 25^oC. The optimum supply voltage is relatively insensitive to temperature variation.

sub-threshold region.

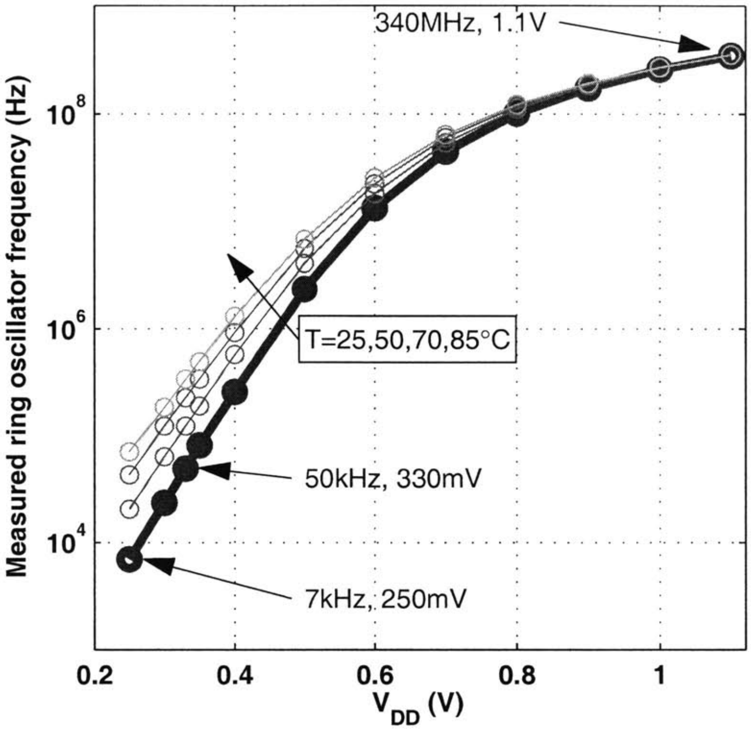


Figure 4-9: Measured ring oscillator frequency versus V_{DD} and temperature.

Figure 4-10 illustrates the savings that LVD provides for the adder block on the test chip when the rate varies. The dotted line shows operation at the highest rate followed by ideal shutdown. The solid line shows the measured energy versus rate for DVS assuming continuous voltage and frequency scaling. Selecting two rates from the curve, 1 and 0.5 in the figure, and operating for the correct fraction of time at each rate results in the dashed line that connects the quantized points, as described previously. A local block with three headers can achieve closer to optimum savings by selecting three rates and then dithering to connect those points on the plot.

While previously reported chip-wide approaches to voltage dithering have largely ignored the overhead energy of their schemes, we have investigated and measured the time and energy overhead of the LVD switching approach. Figure 4-11 shows the test circuit used to measure the delay overhead of LVD. While the adder runs a

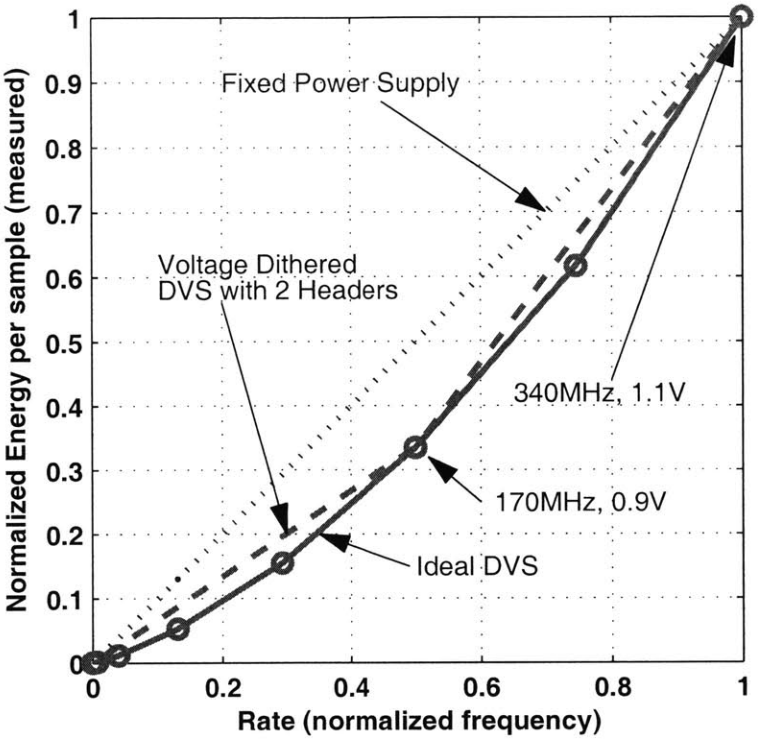


Figure 4-10: Characterized local voltage dithering using measured results for 32-bit Kogge-Stone adder.

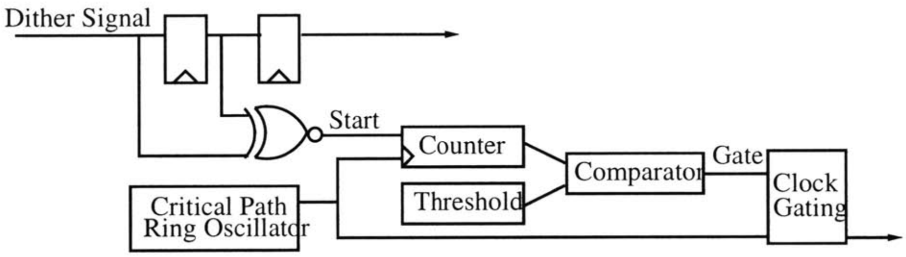


Figure 4-11: Circuit for measuring timing overhead of LVD that gates the clock at a V_{DD} transition for a given number of cycles. The duration of this clock gating is decreased until the circuit fails.

long accumulation, V_{DD} dithers to and from the higher rate. The oscilloscope plot in Figure 4-12 shows the divided ring oscillator output and the signal that selects the supply voltage (dither) for a dithering cycle between full and 0.5 rate. When the headers toggle V_{DD}, a counter gates the clock for a specified number of cycles to ensure settling at the new voltage. Checking the accumulated value verifies correct operation for every cycle. Measurements showed that the correct value was accumulated even with only 1/2 cycle (minimum possible using the test circuit) of clock gating for V_{DDL} above 0.6V, which corresponds to a rate of 0.04. Thus, even conservative settling times for this LVD implementation are on the order of a few cycles. This measurement confirms that LVD can adjust to fast changes in the workload of the local blocks.

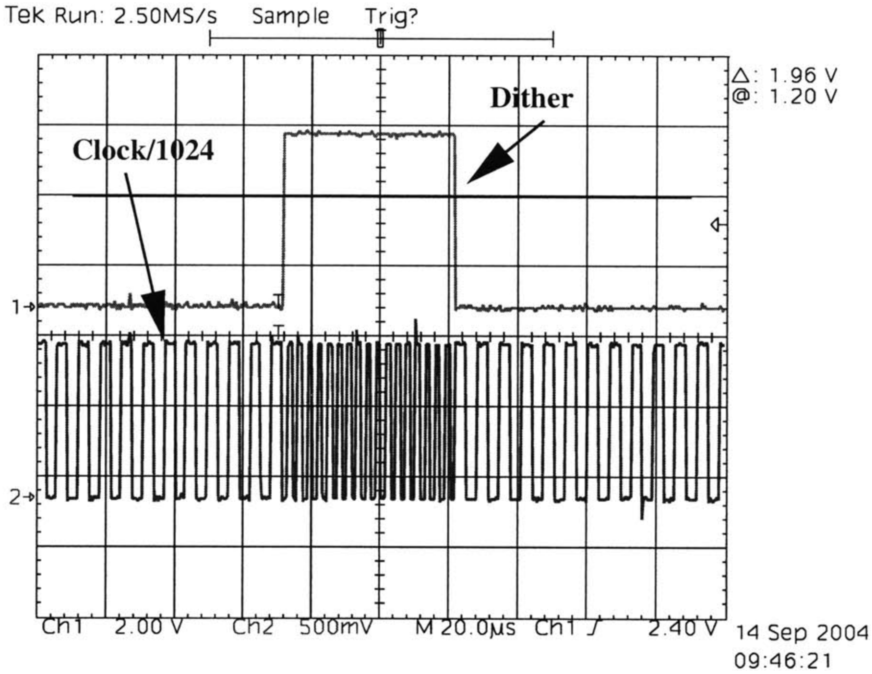


Figure 4-12: Oscilloscope plot showing the system clock while dithering between rate 0.5 (170MHz) and rate 1 (340MHz). Measurements show correct accumulation at both transitions even no clock gating (see Figure 4-11).

In addition to timing overhead, there is energy overhead associated with the LVD approach. The buffer network and control circuits that drive the header switches consume energy every time they toggle the header switches to select a new supply voltage for the adder circuit. We can relate this overhead switching energy to the active switching energy of the adder block to determine its impact on overall energy savings from the LVD approach. To this end, we normalize the effective overhead switched capacitance of the control and buffer circuits, $C_{OVERHEAD}$, to the effective switched capacitance of the adder. The expression in Equation (4.2) shows the relation that must hold true in order for LVD to provide energy savings for a given transition.

$$NV_{DDH}^2 \geq NV_{DDL}^2 + C_{OVERHEAD}V_{DDH}^2 \tag{4.2}$$

Solving (4.2) for N gives the number of cycles that must occur at V_{DDL} in order to make switching to V_{DDL} worthwhile for saving energy, as shown in (4.3).

$$N \geq \frac{C_{OVERHEAD}V_{DDH}^2}{V_{DDH}^2 - V_{DDL}^2} \tag{4.3}$$

Measurements of the test chip show that $C_{OVERHEAD} = 3.7$ for the adder, so N is only 12 for the adder block with V_{DDH}=1.1V and V_{DDL}=0.9 (rate=0.5). Since the control circuits on the test chip are relatively simple, the overhead energy for more complicated control schemes, such as those that calculate the effective workload, has the effect of increasing N.

4.3 UDVS System Considerations

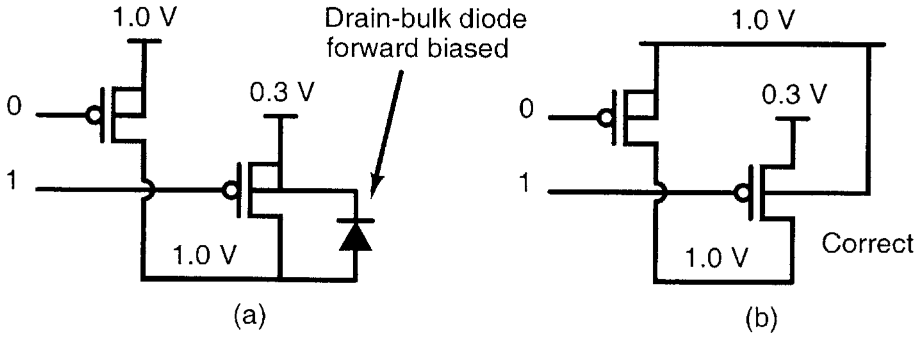


Figure 4-13: For UDVS, the bulk connections of the pMOS header switches have to connect to the highest supply voltage.

The discussion to this point has assumed that the varying rate remains above roughly a few percent. As previously mentioned, sub-threshold operation has proven to minimize energy for low performance applications. While scaling to sub-threshold is rarely advantageous for full processors [77], local blocks or special applications that require brief periods of high performance spend significant amounts of time operating at effective rates that are orders of magnitude below one. Examples of these applications include micro-sensor nodes, medical devices, wake-up circuitry for processors, and local blocks on active processors. When performance is non-critical, energy is minimized by operating at the minimum energy point that occurs because of increased leakage energy at low frequency and then shutting down if there is more timing slack.

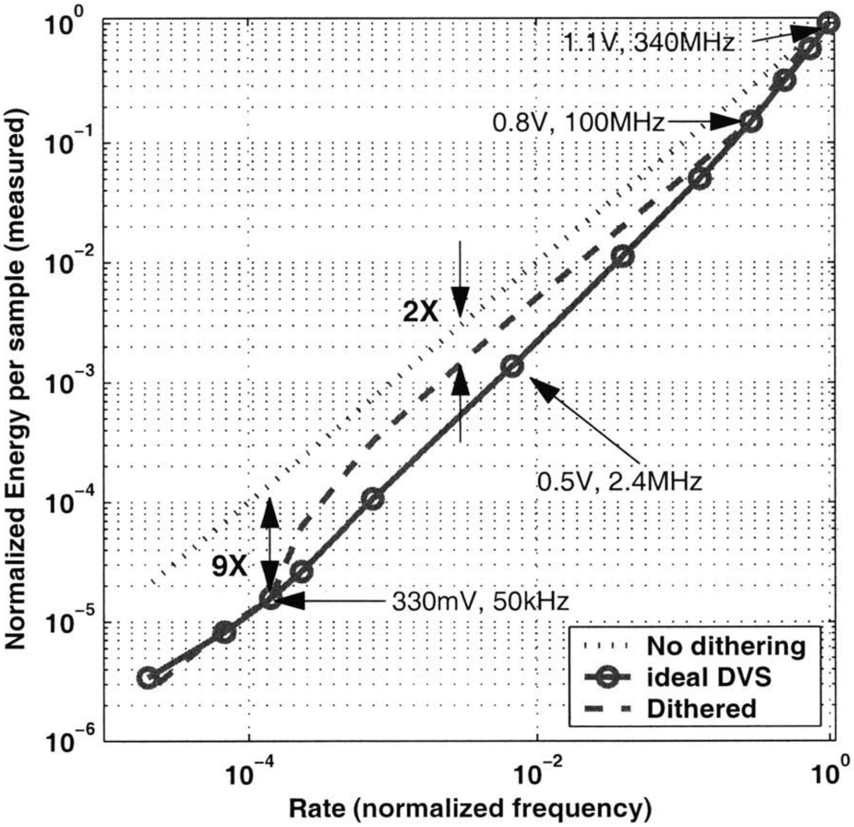


Figure 4-14: Ultra-Dynamic Voltage Scaling (UDVS) using two headers with one variable dc-dc converter or using three headers (c.f. Figure 4-17).

Since LVD works well for high speed operation and operating at the minimum energy point is optimal for low performance situations, we propose ultra-dynamic voltage scaling (UDVS) using local power switches [87]. This approach uses local headers to perform LVD when high performance is necessary and selects a low voltage for sub-threshold operation at the minimum energy point whenever performance is not critical. As with LVD, all of the headers for a given block can turn off when the block is idle to conserve standby power using power gating. Since the power switches in the UDVS approach connect to different voltages that can differ substantially, they must be configured carefully to prevent forward biasing the junction diodes. Figure 4-13 shows that connecting the bulk terminal of the header transistors to the source can forward bias the drain-bulk diode when the V_{DDL} switch is off. One solution to this problem is to tie the bulk of all of the header switches to the highest supply voltage as in Figure 4-13(b).

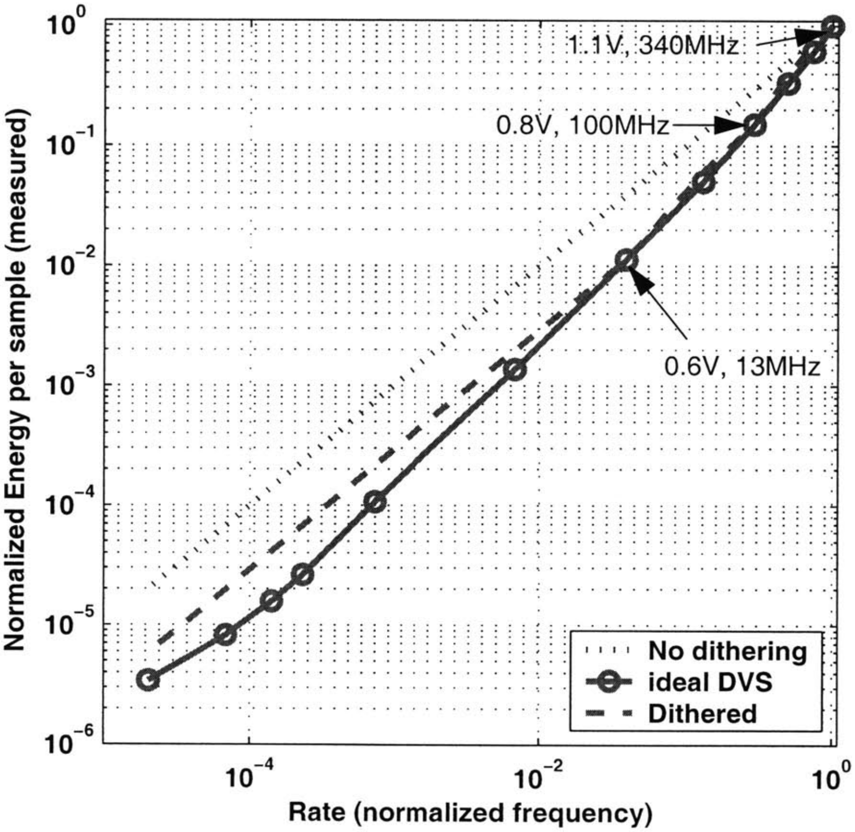


Figure 4-15: Different choice of dithered voltages for closer fit over the higher range of V_{DD}.

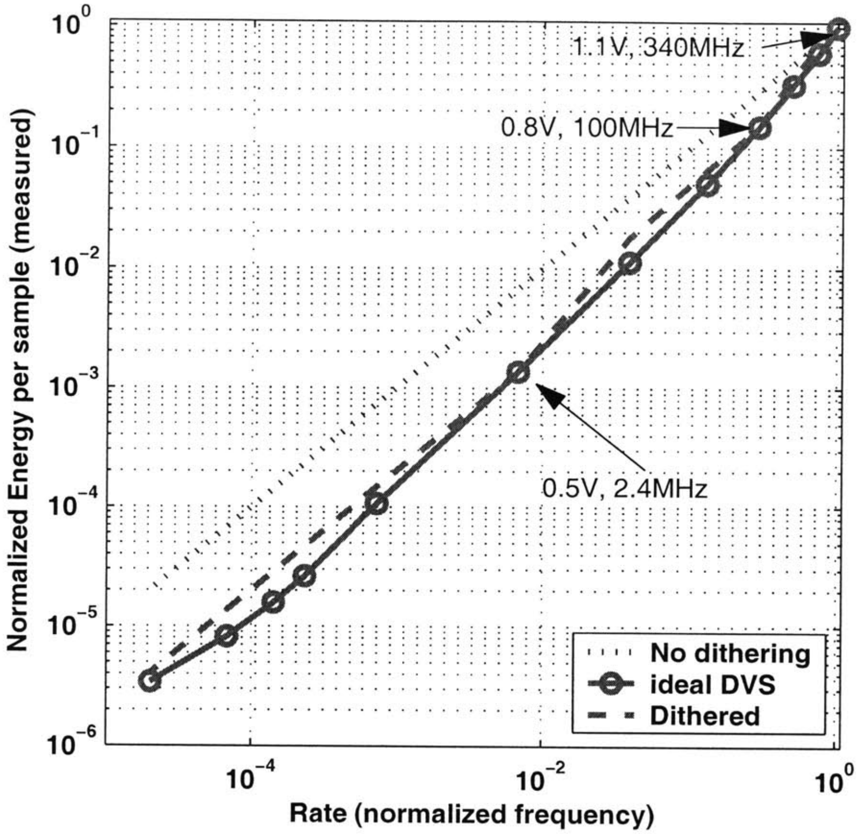


Figure 4-16: Different choice of dithered voltages for closer fit over the entire range of V_{DD}.

Figure 4-14 provides one example of measured UDVS characteristics for the adder. In this example, dithered voltages are chosen at 1.1V, 0.8V, and 0.33V, which is the optimum voltage for minimum energy. When the adder block is performing operations with no timing deadline, it functions at the minimum energy point at 50kHz and saves 9X the energy versus the ideal shutdown scenario. When performance becomes important, the adder dithers between 1.1V and 0.8V within 30% of the optimal energy consumption while adjusting for variations in the rate above 0.1. It was shown in [93] that significant extra savings are available if the selected dithered rates match to the prominent average rates in the data. This brings the dithered curve closer to the optimum DVS curve for the common cases. Figure 4-15 and Figure 4-16 show two additional examples in which the supply voltages are chosen for different scenarios.

For a system whose rate requirements vary evenly over the full range, the voltage choices in Figure 4-15 provide a better match to the ideal energy profile, but the minimum energy per operation is not achievable. If performance constraints prevent a system from ever operating at the minimum energy point, the supply voltage can be adjusted to higher voltages to achieve near optimal energy operation over the range of higher rates (Figure 4-16).

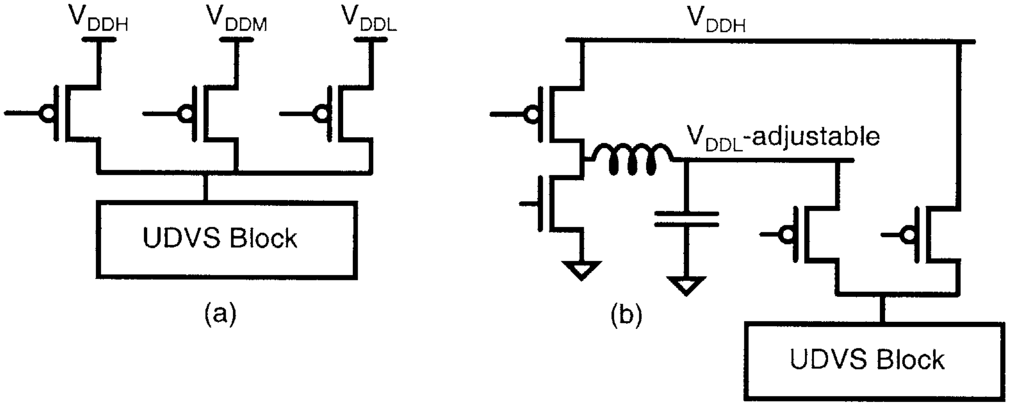


Figure 4-17: Options for UDVS headers at the system level.

Figure 4-17 shows two options for implementing the power supplies and headers in a UDVS system. The straightforward option is to distribute three supply voltages around the chip and to use three header switches at each block. The voltages V_{DDH}, V_{DDM}, and V_{DDL} can be selected based on the system workload statistics as we described above. The only advantage to using more than three power supplies is to pin the UDVS energy profile to the ideal variable supply profile in more places. The right-hand diagram in Figure 4-17 offers a second option for implementing UDVS. When transitions between high and low performance mode are infrequent and the amount of time in between transitions is long, two header switches may be paired with one adjustable dc/dc converter for the same functionality. For example, during high speed operation, the headers dither between 1.1V and 0.8V. When the rare transition to low speed occurs, the dc/dc converter switches V_{DDL} to 0.35V so that all of the blocks can operate near their minimum energy points.

For applications where some blocks operate in sub-threshold while others are at higher voltages, special interfacing circuits are required at the low voltage region to high voltage region interface. The type of level converters to be used will depend on how the block interfaces to surrounding blocks. Ample previous work on level converter circuits offers many choices for implementing the required interfaces. In a full UDVS system with multiple blocks, each block has its own header devices so that it can voltage dither based on its individual workload. Communication among blocks occurs along a bus, which might be asynchronous to account for different operating frequencies, and level converters interface to the bus as needed.

4.4 Summary and Conclusions

Local voltage dithering provides a flexible approach for saving energy in fixed throughput systems. When saving energy becomes more important than meeting a performance constraint, operating at the minimum energy point in sub-threshold offers the best option. For many applications, sub-threshold operation will only make sense when the rest of the chip is off (i.e. wake-up circuits) or when the entire chip is in sub-threshold (i.e. micro-sensor node). This section proposes ultra-dynamic voltage scaling, which combines dithering at high performance and minimum energy operation for low performance scenarios. A 90nm bulk CMOS test chip, shown in Figure 4-4, verifies the energy savings achievable by LVD and UDVS. This flexible approach allows scaling across the full range of supply voltage and frequency in an energy efficient fashion.

Chapter 5

Sub-threshold SRAM

As described in the preceding chapters, sub-threshold digital circuit design has emerged as a low energy solution for applications with strict energy constraints. Analysis of sub-threshold designs has focused on logic circuits (e.g. [4]). The large fraction of chip area often devoted to Static Random Access Memory (SRAM) makes low power SRAM design very important as well. For system integration, SRAM must become capable of operating at sub-threshold voltages that are compatible with sub-threshold combinational logic. Scaling supply voltage for SRAMs has the additional benefit of decreasing their leakage power and active energy. SRAMs comprise a significant percentage of the total area for many digital chips as well as the total power [7][94]. For this reason, SRAM leakage can dominate the total leakage of the chip, and large switched capacitances in the bitlines and wordlines makes SRAM accesses costly in terms of energy. Recent low power memories show a trend of lower voltages with some designs holding state on the edge of the sub-threshold region (e.g. [8]). This scaling promises to continue, leading to sub-threshold storage modes and even sub-threshold operation for SRAMs operating in tandem with sub-threshold logic.

Before examining the details of SRAM operation in sub-threshold, we need to understand what operating scenarios are possible for memory at a higher level. We can define three broad regions of operation for an SRAM as shown in Table 5.1. Proper system-level design of an SRAM to minimize total power and energy consumption requires an understanding of these three modes.

Table 5.1: System-level modes of operation for SRAM.

Operation Mode	Must Keep Data	Read/Write Accesses Allowed
Shutdown	No	No
Idle (Standby)	Yes	No
Active	Yes	Yes

The SRAM provides no functionality in the shutdown state, since any data it previously held does not need to be retained. Consequently, this state provides the lowest possible power consumption for an SRAM. Many applications, such as processors in a PC, may not allow the shutdown state so long as the chip is powered because data needs to be retained. Low power applications such as portable devices or micro-sensor nodes, on the other hand, could easily incorporate SRAM blocks that shutdown. For example, hardware accelerators on a micro-sensor node do not need to retain data in their local memories when they are inactive. Since data retention during shutdown is unimportant, minimizing total memory energy during a shutdown period is equivalent to minimizing leakage power. Any overhead energy expended when entering and exiting shutdown mode places a constraint on when to enter shutdown (i.e., a break-even time), but it does not impact the minimum energy solution once the memory has already entered shutdown. There are many techniques in the literature that can contribute to lowering leakage power for SRAMs including power gating and body biasing (see Section 5.1). For shutdown, V_{DD} for the SRAM can theoretically be reduced to zero.

The SRAM must retain data during the idle state, but it does not need to provide read or write accesses to that data. For a given length of idle time, the total energy consumed by the SRAM equals the leakage energy. As with the shutdown case, leakage energy is minimized when the leakage power is minimized. Thus, the supply voltage can be reduced to lower leakage power, but it must remain above some minimum value to ensure that the bitcells retain their state [95][96]. Other leakage reduction mechanisms are also available, some of which we describe in Section 5.1.

During active operation, the SRAM provides both read and write accesses to the bitcells. Total energy during active operation consists of both an active component and a leakage component. The best approach to minimize the total energy in this case will depend on the size and architecture of the memory and on the specific techniques employed for reducing power. In many cases, idle and active operation are constrained to the same voltage because multiple V_{DD} values are not available. In the case of an SRAM that can never lose its data (i.e., cannot enter shutdown mode), the leakage component of total energy becomes a constant. Minimizing total energy becomes a matter of reducing V_{DD} to lower the active energy and of reducing the effective switched capacitance for a memory access through architectural adjustments.

Although the final system-level design for an embedded low power SRAM depends on how it will function using the three modes we have described, it is clear that lowering the supply voltage is advantageous for reducing the total SRAM power and energy consumption in most cases. The difficulties of operating an SRAM in sub-threshold require both circuit and architectural innovations. The benefits of overcoming these problems are significant, however, since low energy SRAM is essential for enabling ultra-low energy systems.

This chapter describes the investigation of an SRAM capable of operating in the sub-threshold region. Although little previous work addresses this problem directly, we first provide a survey of related low power SRAM techniques from the literature in Section 5.1. Then Section 5.2 describes several key problems that prevent traditional six transistor (6T) bitcells from functioning properly in sub-threshold in a 65nm bulk CMOS process. Section 5.3 takes a deeper look at cell stability by showing and modeling the dependence of Static Noise Margin (SNM) on various parameters. Section 5.4 shows the bitcell that we developed to overcome these challenges to sub-threshold functionality, and Section 5.5 provides results from a 256kb SRAM test chip that uses the new bitcell.

5.1 Low Power SRAM Design

This chapter focuses on reducing SRAM leakage power and active energy using primarily the supply voltage knob. There are many other techniques that can be used in conjunction with voltage scaling. A system-ready SRAM will most likely consist of a holistic mixture of techniques that provide additional gains on top of supply voltage scaling. The following discussion of previous work highlights low power design approaches that are pertinent to extreme voltage scaling or that are congenial to combination into a sub-threshold memory system.

5.1.1 SRAM Overview

To provide background for understanding previous low power techniques, this section gives a brief overview of the traditional 6T SRAM bitcell and its operation in the above-threshold voltage region. Despite numerous attempts to improve upon it, the 6T cell has remained the bitcell of choice for SRAM designs because of its relatively wide noise margins [97]. Figure 5-1 shows a schematic for the basic 6T bitcell. This bitcell essentially consists of back-to-back inverters that store the cell state (M_1,M_3,M_4, and M_6 in Figure 5-1) and access transistors for reading and writing (M_2 and M_5 in Figure 5-1).

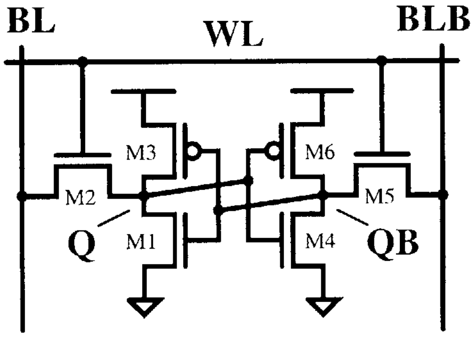


Figure 5-1: Schematic for a standard 6T bitcell.

Figure 5-2 gives a simple example of a memory architecture that uses the 6T SRAM [98]. An array of bitcells is arranged into rows and columns. Each row lies

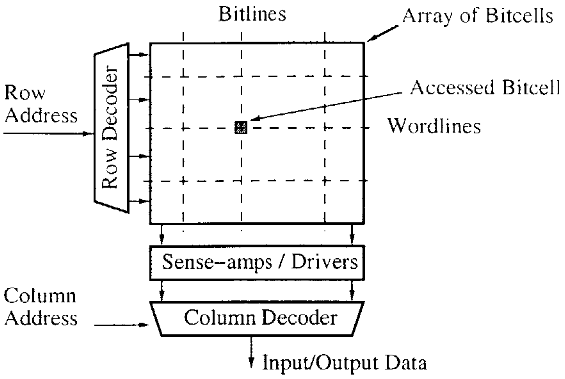


Figure 5-2: Standard architecture for an SRAM using the 6T bitcell.

along a wordline (WL), and each column is associated with a bitline (BL) pair. The memory address is divided into a row address and a column address. Decoder circuits use the applied address to select the correct wordline and bitline pair for a memory access. For a write access, the wordline is asserted (goes to '1') to turn on the access transistors (M_2 and M_5). The bitlines (BL and BLB) are driven to the correct differential value to write into the cell. The bitline driving a '0' will overwrite the data held by the cross-coupled inverters assuming that the bitcell is sized correctly [98]. For a read access, the bitlines are precharged to '1', then the wordline is asserted at the same time as the bitlines are allowed to float. The internal node of the bitcell that holds a '0' will pull its bitline low through the access transistor. Usually, a sense amplifier will detect this differential voltage on the bitlines before it becomes very large and amplify it to full voltage values. Sense amps are used primarily to speed up the read process or to avoid the energy overhead of fully discharging the large capacitance of the bitlines.

When the bitcell is holding data, its wordline is low so M_2 and M_5 are off. In order to hold its data properly, the back-to-back inverters must maintain bi-stable operating points. The best measure of the ability of these inverters to maintain their state is the bitcell's SNM [6]. The SNM is the maximum amount of voltage noise that can be introduced at the outputs of the two inverters such that the cell retains its data. SNM quantifies the amount of voltage noise required at the internal nodes of a bitcell to flip the cell's contents.

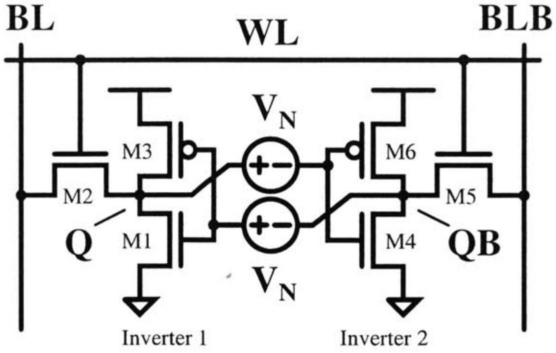


Figure 5-3: Schematic for 6T bitcell showing voltage noise sources for finding SNM [6].

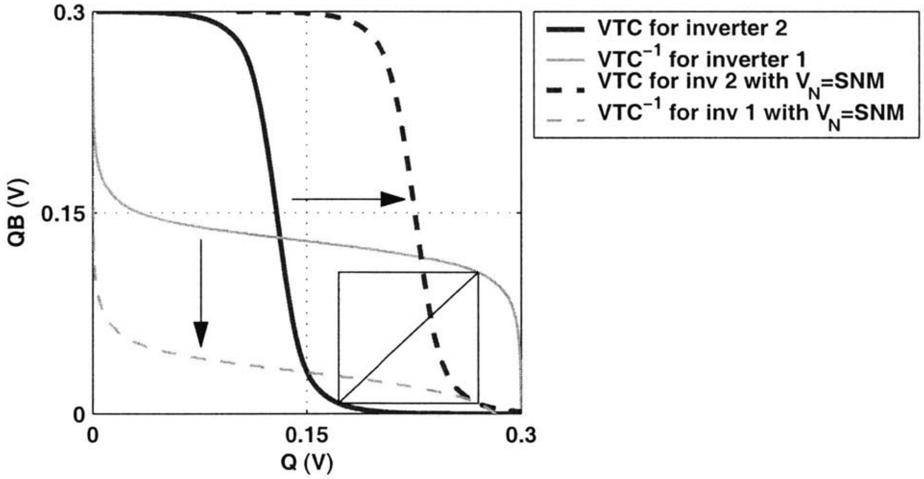


Figure 5-4: The length of the side of the largest embedded square in the butterfly curve is the SNM. When both curves move by more than this amount (e.g. V_N=SNM), then the bitcell is mono-stable, losing its data.

Figure 5-3 shows a conceptual setup for modeling SNM [6]. Noise sources having value V_N are introduced at each of the internal nodes in the bitcell. As V_N increases, the stability of the cell changes. Figure 5-4 shows the most common way of representing the SNM graphically for a bitcell holding data. The figure plots the VTC of Inverter 2 from Figure 5-3 and the inverse VTC from Inverter 1. Figure 5-5 shows the equivalent circuits from the bitcell used to generate the VTCs in the butterfly plot.

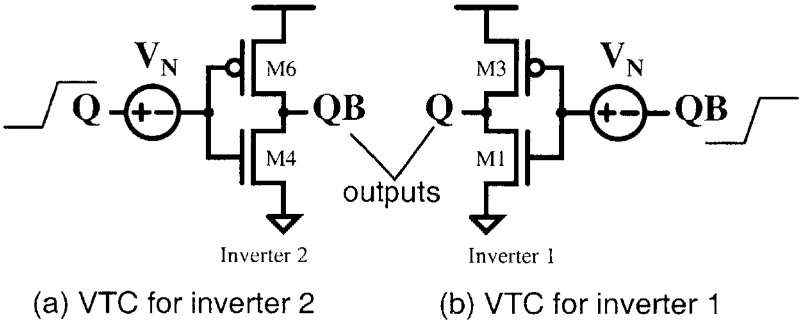


Figure 5-5: Equivalent circuits for 6T bitcell for finding VTCs in the butterfly curve. (a) shows the equivalent circuit for the VTC of inverter 2, where Q is swept and the output is at QB. (b) shows the equivalent circuit for the VTC of inverter 1, where QB is swept and the output is at Q (the butterfly curve shows the inverse of this because Q is on the x-axis).

The resulting two-lobed curve is called a "butterfly curve" and is used to determine the SNM. The SNM is defined as the length of the side of the largest square that can be embedded inside the lobes of the butterfly curve [6]. To understand why this definition holds, consider the case when the value of V_N increases from 0. On the plot, this causes the VTC^{-1} for Inverter 1 in the figure to move downward and the VTC for Inverter 2 to move to the right. As V_N increases, the metastable point moves closer to one of the stable points in the plot (the lower-right point in this example). Once both curves move by the SNM value, the metastable point becomes coincident with one stable point, and the curves meet at only two points. Any further noise flips the cell.

Although the SNM is certainly important during hold, cell stability during active operation represents a more significant limitation to SRAM operation. Specifically, at the onset of a read access, the wordline is '1' and the bitlines are still precharged to '1' as Figure 5-6 illustrates. The internal node of the bitcell that represents a zero

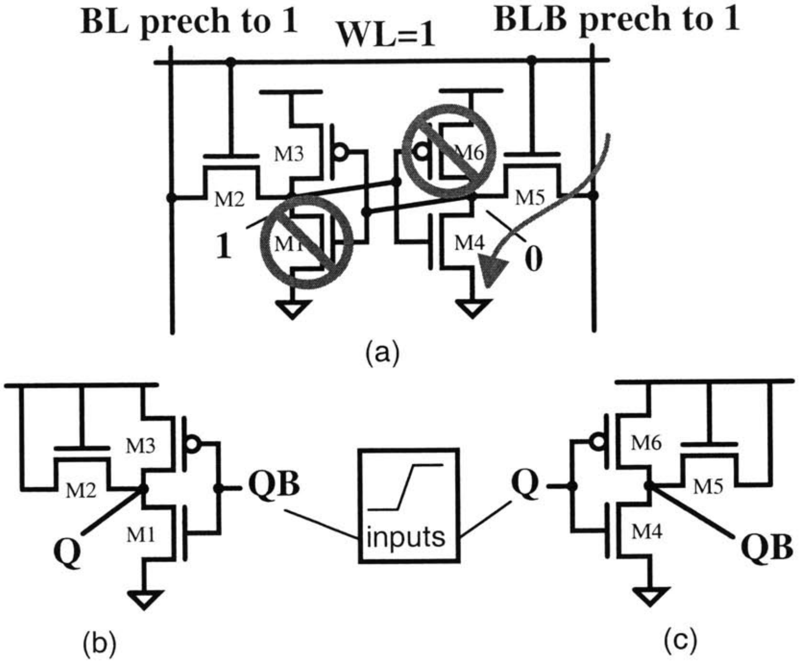


Figure 5-6: Schematic of the 6T bitcell at the onset of a read access (a). WL has just gone high, and both BLs are precharged to V_{DD}. The voltage dividing effect across M_4 and M_5 pulls up node Q_B, which should be 0V, and degrades the SNM. (b) shows the equivalent circuit of the bitcell for plotting the VTC of inverter 1 in the butterfly curve (QB input, Q output). (c) shows the equivalent circuit of the bitcell for plotting the VTC of inverter 2 in the butterfly curve (Q input, QB output).

gets pulled upward through the access transistor due to the voltage dividing effect across the access transistor (M_2,M_5) and drive transistor (M_1,M_4). This increase in voltage severely degrades the SNM during the read operation (read SNM). Figure 5-7 shows example butterfly curves during hold and read that illustrate the degradation in SNM during read. The voltage dividing effect causes the lower half of the VTC for each inverter (when its V_{in} is high) to pull upwards relative to its original position, squashing the lobes of the butterfly curve.

The following discussions of previous work deal with variations to this standard 6T bitcell and architecture. The focus of most of the works we report is to lower power, and many of the techniques can be combined with the sub-threshold SRAM that this thesis presents.

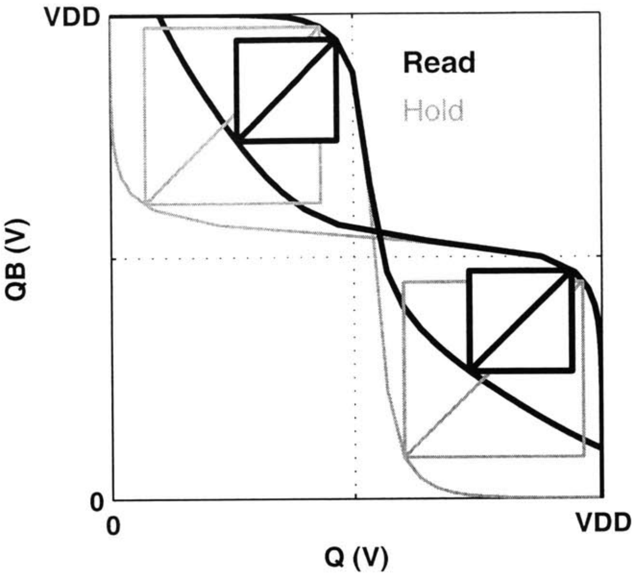


Figure 5-7: Example butterfly curve plots for SNM during hold and read.

5.1.2 Voltage Scaling

Scaling of the supply voltage and other voltages related to SRAM operation has become a popular method for improving on the basic architecture. Table 5.2 shows a sampling of the different methods of voltage scaling applied to SRAM operation.

The significance of SRAM power has produced a trend of memory design aimed at lower voltage operation. Exploiting DVS for SRAM is one motivation for designing a voltage-scalable memory. A 0.18μm 32kB four-way associative cache offers DVS compatability from 120MHz, 1.7mW at 0.65V to 1.04GHz, 530mW at 2V [111]. This memory uses high V_T in the bitcell array and low V_T transistors in the peripheral circuits. Since lowering V_{DD} amplifies the difference in delays between these two V_T regions, the architecture uses dummy bitlines to adapt the timing correctly with scaling. The bitcell itself is lithographically symmetrical to reduce the impact of lithographical mismatch on delay at lower V_{DD}, similar to [112]. Although DVS can provide power reduction for active memories, the more common approach to voltage scaling is to implement it primarily for idle SRAM blocks.

Reducing the voltage in an idle memory array lowers the standby power. Figure 5-

Table 5.2: Voltage scaling approaches used for SRAM.

Voltage	Approach	Source(s)
bitcell V_{DD}	lower in standby	[94][96][99][8][100]
	raise always	[101][102]
	raise for read access	[8][103]
	float or lower for write	[8][104]
	float for read access	[104]
	raise in standby	[105]
bitcell V_{SS}	raise in standby	[105][106][107][108][7][109][100]
	raise or float for write access	[7][110]
	lower for read access	[103]
wordline	negative for standby	[99][104]
WL driver V_{DD}	lower in standby	[7]
well biasing	change with mode	[99][103]
bitline V_{DD}	lower for standby	[106]

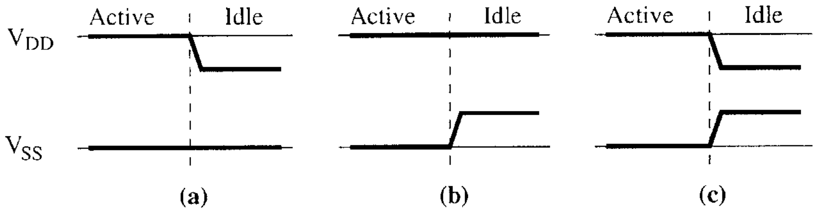


Figure 5-8: General approaches using voltage scaling to lower idle power in SRAM. Lowering V_{DD} (a), raising V_{SS} (b), or both (c).

8 shows three general methods for implementing standby voltage scaling. The power supply itself is reduced in Figure 5-8(a) (e.g. [94][96][99][8]), the ground voltage is increased in Figure 5-8(b) (e.g. [105][106][107][108][7][109]), and both rails are scaled in Figure 5-8(c) (e.g. [100]). For the case where V_{DD} is lowered for idle cells, the minimum voltage for retaining bistability was theorized in [46] and modeled for SRAM in [96]. Implementations of SRAM using lower V_{DD} in standby are available [99] along with software policies to determine when to enter the lower leakage mode [94].

One issue for deeply voltage scaled SRAM is Soft Error Rate (SER). Soft errors occur when an alpha particle or cosmic ray strikes a memory node and disrupts it such that it loses its value. Since the susceptibility of a memory node to a soft error is proportional to the amount of charge stored there, there is a minimum amount of capacitance that should be present in a cell to keep SER acceptably low [113]. Since bitcell storage capacitance decreases with scaling and voltage scaling further reduces the stored charge, SER is a concern for sub-threshold memory. Fortunately, there are methods for taking care of soft errors. Studies of soft errors have shown that multi-cell errors from a single strike only occur in a limited number of cells along a wordline (2 to 3) [114]. Thus, physically interspersing bits from different words can prevent multi-errors from occurring in a single word [114]. The additional application of error correcting codes using parity bits can fix any such errors that occur. A chip that implements these techniques reports excellent success in reducing SER [106].

A common alternative to lowering V_{DD} for reducing standby leakage is to increase the virtual ground node, V_{SS}, as in Figure 5-8(b). By explicitly raising V_{SS} or allowing it to float to higher voltages, transistor V_{DS} reduces and RBB further reduces leakage by increasing device V_Ts. In [7] and [106], V_{SS} is raised explicitly to 0.3V and 0.5V, respectively. In [108], the virtual ground node is allowed to float, and its voltage is limited by a diode-connected transistor.

Scaling the supply rail voltages is often supplemented by other leakage reducing approaches for standby. For example, negatively biasing the wordlines reduces leakage into the bitcells through the access transistors [99][104]. The negative wordlines are combined with V_{DD} lowering and with N-well biasing to match pMOS and nMOS leakage currents in [99]. Further precautions such as nMOS pullups on the bitlines lighten the stress on the negatively driven access transistor gates. The delay associated with charging V_{DD} is estimated to be only 5% of the total read delay [99].

Well-biasing is also used specifically for leakage reduction in standby [115][103]. In [103], a triple-well process allows RBB during standby and Forward Body Bias (FBB) during active operation. The various voltages required for this approach are generated off chip, and thick oxide devices are required in some places to withstand larger than normal gate voltages.

Other memories play more tricks with voltage supplies. In [101][102], the V_{DD} to the bitcells always is boosted relative to the periphery's V_{DD} by 100mV, and the chip works to a periphery voltage of 0.4V. The higher V_{DD} at the cross-coupled inverters improves Read SNM and reduces read delay by strengthening the drive transistors relative to the access transistors. The same effect degrades the ability to write, so the pMOS transistors in the bitcell use higher V_T to allow robust write operation.

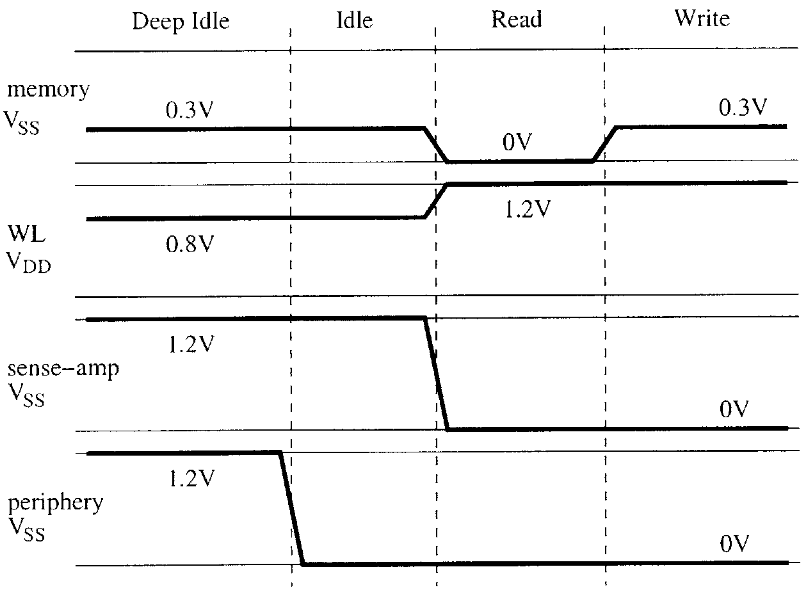


Figure 5-9: Voltage scaling strategies for different operating modes in [7].

A large SRAM in [7] uses three supply voltages (0.3V, 0.8V, and 1.2V) to implement different operational modes as illustrated in Figure 5-9. During deep standby, V_{SS} increases to 0.3V, the wordline V_{DD} drops to 0.8V using a diode-connected FET, and the sense amps and other peripherals are fully power gated by pulling their ground nodes to 1.2V. V_{SS} also is increased to an entire row during write access.

An SRAM proposed in [8] for a CMOS-SOI process floats V_{DD} to an accessed row during both read and write as shown in Figure 5-10. During reading, the rising wordline pulse is intentionally coupled onto the floating V_{DD} of the bitcells by routing the wordline directly alongside the virtual power rail. This improves both the read SNM and speed. During a write access, the floating V_{DD} collapses to make writing easier. A diode stack serves to raise V_{SS} during standby.

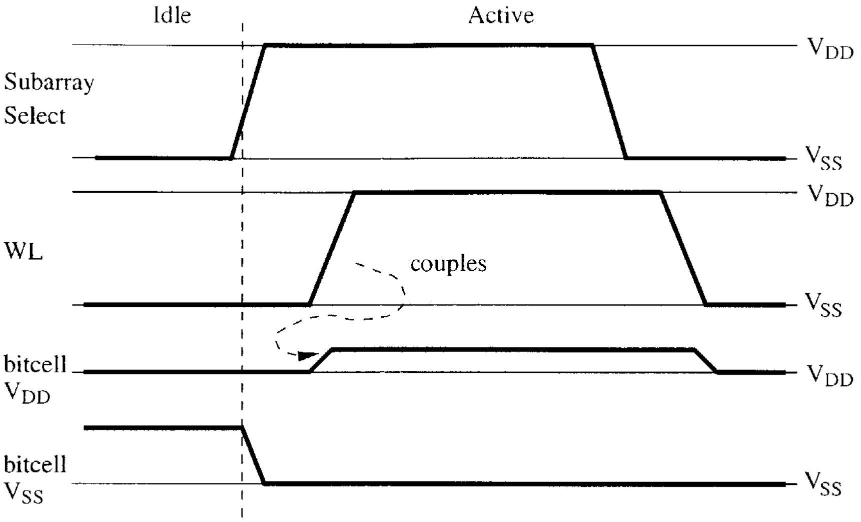


Figure 5-10: Voltage is boosted to accessed cells using capacitive coupling from the wordline in [8].

Clearly, previous efforts have explored many options for voltage scaling. However, none have yet pushed voltage scaling into the sub-threshold region during active operation.

5.1.3 Write Access

Although the write access to an SRAM bitcell historically causes fewer problems than read and occurs less frequently, it has received increased attention lately. This section highlights a class of approaches that adapts well to sub-threshold operation.

During a write access in the normal implementation, there is a ratioed fight between the cross-coupled inverters in the bitcell and the write driver combined with the access transistor. We describe this issue in detail in Section 5.2.1. One approach to remove the ratioed nature of this fight cuts off either one of the supply rails inside the bitcell to weaken the cross-coupled inverters. This approach might simply allow the virtual node to float or might actively drive it to a different value. Examples exist for both a virtual ground node [104][7][7][116] and a virtual power node [8].

One issue with writing is that the power expended to drive the full capacitance of the bitlines full-swing may be unnecessary. Several approaches have emerged to execute the write with lower swing bitlines. One example that extends the concept described above adds an extra nMOS in series with the V_{SS} rail in the 6T bitcell [110]. During a write, this nFET turns off so that the internal V_{SS} node floats. Then the write drivers assert a small differential voltage ($V_{DD}/6$) on the bitlines. Finally, the tail nFET turns back on and the feedback inside the cell amplifies the bitline differential and drives it to full swing. All of the cells along the wordline are written at once, and they can share the V_{SS}-gating switch. This approach reportedly reduces power by 90% [110].

Since sizing is a less effective knob in sub-threshold, the ratioed nature of traditional write access creates a problem. The technique of weakening the cross-coupled inverters by gating their supply voltage or ground node removes this problem and seems applicable to sub-threshold memories.

5.1.4 Data Sensing for SRAM Read

Traditional SRAM designs use sense amplifiers to detect small voltage differentials on the bitlines and to amplify them to full-swing. Changes in technology resulting from DSM scaling are reducing the improvements achieved by differential sensing. As technologies scale, the contribution of the interconnect to bitline capacitance begins to dominate the diffusion capacitance of the cells [117]. This effect means that the differential signal development does not speed up much with smaller technologies. Also, large numbers of cells on the bitline contribute leakage currents that oppose the desired discharge, slowing the differential signal development even further. As a result, differential SRAMs use fewer cells on a bitline pair by about 2x every 2 generations with 64 or fewer rows per bitline in $0.13\mu m$ technology [118]. Device sizes in the sense amplifier must increase with technology scaling to account for variations, so the overall area and delay efficiency of the sense amplifier approach degrades [117]. The comparison of the two approaches in [117] for a 512-row by four-column block shows that small signal reading is more area efficient and faster for technologies larger than 90nm. However, at 90nm and below, the large signal approach maintains roughly constant area efficiency while its delay reduces with technology in proportion to the delay of generic logic. In contrast, the small signal approach does not decrease the delay much as technology scales, and its area efficiency drops significantly. Below 90nm, large signal sensing offers faster read times for similar area efficiency [117]. For these reasons, single-ended sensing offers a competitive alternative in deep sub-micron technologies to differential sensing in terms of area efficiency and delay [118][117]. This full-swing, single-ended approach to reading also appears in [119] and [120]. If full-swing read shows promise for DSM above-threshold SRAMs, then it is worth exploring for sub-threshold operation as well. Scaling full-swing read into sub-threshold promises to be much simpler than designing a sense-amplifier that is capable of operating across both regions.

Bitline leakage has become a significant problem for both single-ended and differential sensing read circuits. Bitline leakage refers simply to leakage currents from the bitline that primarily go into unaccessed bitcells. For an unaccessed bank, bitline leakage increases the standby power of the memory. During access, bitline leakage creates several problems. At the onset of a read access, bitline leakage can oppose the on-current of the accessed cell and increase delays for the read operation or even create reading errors. The worst-case bitline leakage occurs when all of the rows in a bank hold the complement of the value in the cell that is addressed. Thus, on the bitline that should be '1', leakage through the access transistors into unaccessed cells acts to pull down the bitline voltage. If this leakage is large enough, then the '1' bitline can initially drop its voltage even faster than the bitline which properly should be '0'. Thus, it takes much longer for the bitline differential to develop such that the sense amplifier can turn on. In the case of single-ended read, the degraded '1' value must be distinguishable from its counterpart in steady-state for functional correctness.

Previous work addresses bitline leakage in several ways. First, as we mentioned above, the number of rows on a bitline is decreasing with technology scaling to reduce the impact of bitline leakage. Fine-grained partitioning of larger bitlines into hierarchical stages with the lowest level touching only a limited number of cells has the same purpose (e.g. [119]). For reducing bitline leakage only during standby, previous work has applied RBB to the access transistors [103] and lowered the bitline voltage relative to V_{DD} [106]. Other approaches seek to reduce bitline leakage during an access through circuit modifications. Reducing the voltage of the unaccessed wordlines to a negative value has become popular (e.g. [99]). This creates a negative V_{GS} for the *off* access transistors and reduces their sub-threshold leakage exponentially, but it requires the overhead of creating a negative voltage. It can also potentially cause problems by over-stressing the gate oxides in the access transistors. A different approach boosts the wordline voltage relative to the bitcell V_{DD} during an access (e.g. [121][122]). This strengthens the drive current of the accessed bitcell and speeds up the differential development on the bitlines. By strengthening the access transistor relative to the drive nMOS inside the cell, however, this approach degrades Read SNM.

One more complicated approach to bitline leakage involves measuring the actual leakage current and then compensating for it directly [123]. During the precharge stage, the leakage on the bitlines is allowed to discharge the voltage at certain bias nodes, effectively storing a measure of the leakage on each bitline. During the access stage, the stored values are used to inject a current to each bitline that compensates the leakage. This allows the read access to continue without seeing the negative effects of the bitline leakage. This approach has the disadvantage of adding delay overhead for measuring and injecting the compensation currents.

A simpler compensation approach equalizes bitline leakage onto the two bitlines to first order by adding two transistors to the 6T bitcell. The new 8T bitcell connects its extra transistors from nodes Q and QB to the *opposite* bitlines from the original access transistors and sets their gates to '0' [124], as shown in Figure 5-11. This scheme essentially forces the worst-case amount of leakage onto each bitline but ensures that this leakage is the same for both bitlines. This equalization of current reduces the bitline differential development time by around 80% by removing the need for the accessed transistor to first overcome the voltage offset created by leakage [124]. Interestingly, although the bitcell itself is 40% larger, the resulting cache has

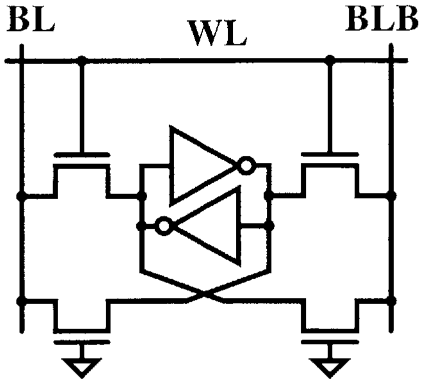


Figure 5-11: An 8T bitcell in [8] equalizes leakage to the bitlines by forcing the worst-case leakage current to both.

6% smaller area due to the integration of 256 rows per bitline rather than only 16 rows [124].

5.1.5 Address Decoding

The address decoder uses the address to determine which memory cell requires accessing and asserts the appropriate wordline and bitlines. Since decoding lies on the critical path for a memory access, the primary concern has been to decrease the delay of these circuits. However, several previous papers have examined address decoder circuits with power in mind. The input capacitance of local and global word drivers is a large contributor to power and delay. Setting the input stage of the word drivers to minimum size and then independently sizing up that chain with fanout of four in each successive stage to drive the actual wordline load achieves near the optimum power-delay curve [125]. Architectures using divided wordlines to reduce power and delay by lowering the capacitance are also popular (e.g. [126]).

The actual circuits used to implement the decoders vary from simple static gates to dynamic logic to pulsed and self resetting logic. For decoders using static logic, two-input NAND gates followed by inverters offer the minimum delay solution [127]. In [128], two-input gates are proven to be best for the decoding stages that follow the pre-decoder, sizing for a fanout of four is optimum even when considering RC interconnect parasitics.

Some fancier options are also available. Since only the 0 → 1 transition needs to be fast for the wordline, the assert and de-assert paths can be decoupled by using a reset pullup in the static NAND gates. This permits skewing the sizes of the transistors on the assert path to make the gates faster for the critical transition, and it saves power by reducing the sizes of the de-assert transistors [125]. Converting the decoded signal to a pulse and adding extra pre-decode logic enables both the capacitive load and the number of stages along the critical path to decrease in [129]. Pulsed circuits coupled with self-resetting CMOS logic are combined with redundant wordlines in [130] to decrease delay.

5.1.6 Alternative Bitcells

This section gives a sampling of the types of changes that have been proposed for the standard 6T bitcell. These alternative bitcells divide roughly into two classes: cells that trade off some metric (e.g. SNM, speed) for reduced area and cells that trade off area to improve some characteristic of the memory (e.g. faster, lower power).

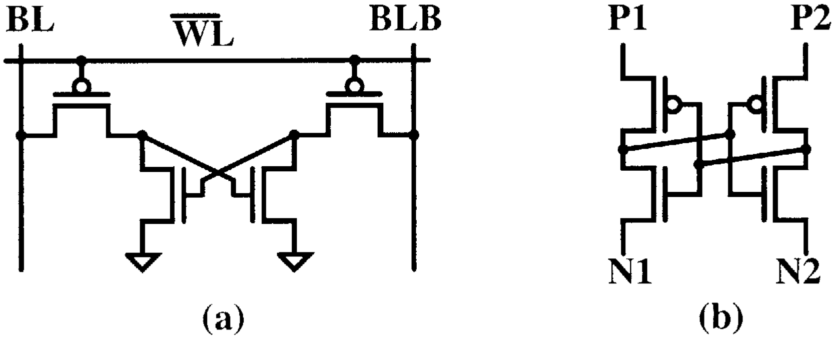


Figure 5-12: Four transistor bitcells. Leakage through low V_T pMOS access transistors provide the load for the cell in (a) [9], and manipulating the sources of the transistors in (b) provides access to the bitcell [10].

Smaller bitcells that retain static behavior reduce SRAM cost by reducing the area. Figure 5-12(a) shows one example of a 4T bitcell uses cross-coupled nMOS pull-downs with no explicit load and two pMOS access transistors [9]. Write and read accesses to this cell occur similarly to 6T, but the cell depends on a relationship between the off-currents of the pMOS and nMOS transistors to ensure that it does not need refreshing during hold. Specifically, $I_{offP} > I_{offN}$ to provide a load for the cross-coupled nMOS transistors. Lower V_T pMOS transistors are used to maintain this relationship [9]. Figure 5-12(b) shows a second type of 4T bitcell consisting of cross-coupled inverters only [10][131]. The sources of all four transistors are used for reading and writing. To write, the supply voltage to the cross-coupled pair is lowered and then the source of one nMOS or the other is pulled high to flip the cell. To read, the source of one nMOS is driven to a higher voltage and a current sensor examines the change in current (if any) that occurs in the opposite inverter. Although the area of this cell is lower, it suffers from complicated peripherals, sensitivity to voltage levels, and difficulty integrating cells into large arrays.

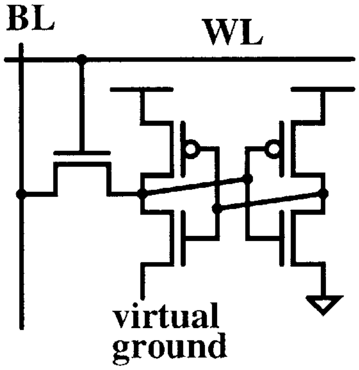


Figure 5-13: Single ended read and write access allow the omission of one access transistor in this 5T bitcell. Single-ended write is enabled by allowing the labelled virtual ground node to float during write access [11].

Figure 5-13 shows a 5T bitcell that looks like a 6T bitcell minus one access transistor and one bitline [11]. Both read and write accesses are single-ended and occur from the single bitline. The source of the nMOS driving the readable node connects to a virtual rail. During write, this virtual ground node floats in a fashion similar to the collapsing supply methods we discussed in Section 5.1.3. A second 5T cell in

0.18μm CMOS looks like the 5T bitcell in [11] but without the virtual ground node [132]. The cell is sized so that writing occurs in a single-ended fashion via the single bitline. Since driving the bitline to V_{DD} writes a '1', the bitline must precharge to an intermediate value for read accesses. In this implementation, the optimal value is 600mV. A sense-amplifier compares this mid-rail value to the accessed bitline after the wordline goes to '1' and the accessed bitcell has pulled its bitline either higher or lower than 600mV. This bitcell reduces area and bitline leakage, but the Read SNM degrades by about 50% [132].

Some variations on the 6T bitcell preserve its basic structure. For example, the 'asymmetrical cell' in [119][120] is precisely a 6T bitcell, but it uses asymmetrical sizing. Write operation uses both bitlines in the normal fashion, but the read access is single-ended and full swing using just one bitline. Since the read occurs on only one side of the cell, asymmetrical sizing can shift the VTCs of the inverters in the cell to increase the Read SNM on that side, improving the bitcell's capacity for low voltage operation. This SRAM also uses divided wordlines and split bitlines. Other variations on the 6T bitcell explore combinations of transistors having either high V_T or low V_T [101][102][118]. An analysis of several dual V_T bitcell designs asserts that the best delay, power tradeoff occurs for a bitcell with minimum length, low V_T access transistors, high V_T cross-coupled inverters, and a low V_T periphery [118].

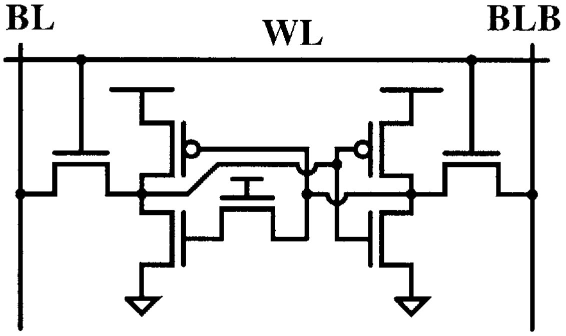


Figure 5-14: A 7T bitcell that takes advantage of the predominance of zeros at the bit level to decrease gate leakage [12].

Some larger bitcells utilize extra area to improve some characteristic of the SRAM relative to a 6T design. Figure 5-14 shows an asymmetric 7T cell that inserts an extra transistor in series with the gate of one of the nMOS drive transistors [12]. This bitcell takes advantage of the predominance of zeros at the bit level in most scenarios to decrease the gate leakage of the bitcell at the cost of area and degraded speed for reading on one side. Other examples of larger cells include the 7T bitcell discussed in Section 5.1.3 that uses an additional cell to reduce write power [110] and the 8T bitcell described in Section 5.1.4 that equalizes bitline leakage to improve read access speed [124].

5.1.7 Embedded SRAM versus Embedded DRAM

Trends in micro-processor design and ASIC design for embedded systems indicate that memory consumes increasingly more die area. Integrated level-1 caches on micro-processors significantly increase speed relative to off chip caches. Even level-2 caches are starting to appear on the same die as the processor. For reasons of cost and speed, these integrated memories traditionally are implemented as SRAM. The huge advantage to SRAM for embedded memory is its direct compatability to logic process technologies. Improvements to process technology coupled with the need for integrating more memory onto a single die fuels a new movement to integrate Dynamic Random Access Memory (DRAM) onto the same die as logic. DRAM's higher density allows more memory to fit on the die, and it typically outperforms SRAM in speed. In order for embedded DRAM to work, it clearly must be manufacturable in the same process as the logic. This can occur in one of three ways. First, adding metal layers to a commodity DRAM process allows logic to be implemented on a classically DRAM die. The logic in this case will have lousy performance. Secondly, DRAM can be added to a logic process by adding processing steps. This approach increases the cost of production and typically results in a larger DRAM cell than the commodity product. The larger cell comes from the different method of aligning salicided gates for logic, which prevents the use of commodity DRAM's self-aligned contact step [133]. The current method of choice is adding steps to a logic process to allow the inclusion of DRAM. The number and cost of the extra steps can be significant, because there is a large disparity between a typical logic process and a commodity DRAM process [134].

One example of an embedded DRAM (eDRAM) technology for 0.18μm CMOS adds a trench capacitor module and four extra masks to the standard logic process [135]. Implementing the eDRAM cell as a trench capacitor is easier than using the alternative of a stacked capacitor (STC) cell, because the STC cell has a significantly larger height than the logic circuits, making integration difficult [133]. As expected, the eDRAM bitcell in this 0.18μm eDRAM process is larger than its commodity counterpart. This eDRAM process has a retention time of greater than 256ms at 85^{o}C [135].

A 16Mb eDRAM operates at 0.6V and 205MHz, consuming 39mW [136]. Although this paper does not report the extra processing steps required, it does state that the eDRAM uses dual V_T transistors, FBB and RBB, triple wells, and two extra external voltage supplies in addition to the bitcell V_{DD}. The FBB and RBB improve speed and lower standby power, respectively. The storage capacitor in the bitcell is 40fF, and the memory has a 128ms refresh time at 0.55V and 85^{o}C [136].

A 10Mb eDRAM differs from the previous examples in that it is built in unmodified 150nm CMOS logic process [137]. The bitcell consists of two pMOS transistors. One pMOS acts as a planar MOS storage capacitor of 5fF. The gate voltage of this pMOS stays at negative 100mV to ensure that the channel remains inverted despite the stored value. The access transistor has a longer than minimum gate length and uses wordlines at negative 150mV when it is unaccessed to reduce leakage current and extend refresh time. This bitcell gives $\sim$ 3X reduction in bit density relative to SRAM. Due to the smaller storage capacitance and unmodified process, the refresh time is shortened to 100μs at 110^{o}C. The refresh power of 42mW is similar to the leakage power of a 6T SRAM of the same area, but the bit density is 3X better for the DRAM. The chip needs 6% redundancy to account for bit errors [137].

Integrated DRAMs tend to cost around 25% more than commodity DRAMs [134]. For this reason, embedded DRAM will have a higher per-chip cost, but it might still make sense when evaluated at the system level. In applications where eDRAM replaces some off-chip storage or where high bandwidths are necessary, it can cut total costs. As eDRAM becomes a more developed field, the number of extra steps added to a logic process will reduce to just the storage node and transfer gate [134].

The ideal embedded DRAM provides compatability with a logic process and requires as few extra processing steps as possible. Due to its smaller size, embedded DRAM offers higher memory density, faster speed, and comparable power to SRAM. The increased cost of extra processing steps and added complexity is the primary disadvantage of eDRAM. Additionally, eDRAM designs cannot easily port to new technologies the way that logic does because they follow non-scalable rules [134]. Embedded DRAM can be optimized for high speed, low-cost, and low-voltage by tailoring the design and processing steps [133]. With this flexibility and the improvement of eDRAM processes, eDRAM is sure to gain a share of the embedded memory market. However, the low cost, reduced complexity, and scalability of embedded SRAM ensure that it will continue to be the right choice for embedded memory for a wide range of different chips.

5.2 6-Transistor SRAM Bitcell in Sub-threshold

Traditional 6T SRAMs face many challenges in DSM technologies, and low V_{DD} operation exacerbates the problems. This section describes key obstacles to sub-threshold SRAM operation for a 65nm process. Predictions in [138] suggest that process variations will limit standard 90nm SRAMs to around 0.7V operation for two primary reasons: degradation of SNM and reduced write margin. Variations in the bitcell transistors caused by phenomena such as global process variation, random doping mismatch, and temperature changes degrade the SNM. The impact of random local variation increases for DSM devices because of the smaller transistor channel area. In the sub-threshold region, variations in the threshold voltage impact delay and current exponentially. Previous work has measured the minimum voltage for retaining SRAM state during idle mode at several hundred millivolts for a 90nm memory [96]. Interestingly, the sensitivity of the SNM to threshold voltage mismatch actually decreases in the sub-threshold region [8], but the lower V_{DD} decreases the absolute value of SNM. Likewise, write access into the bitcell becomes less certain at lower supply voltages. Since standard write operation depends on a carefully balanced ratio of currents, processing variation makes this ratio difficult to maintain as V_{DD} decreases, leading to errors during write access.

These practical problems associated with low voltage operation for SRAMs limit the traditional 6T bitcell and architecture to higher voltage, above-threshold operation. Reports in the literature of 65nm SRAMs confirm this voltage barrier. A 65nm SRAM built in a dynamic-double-gate SOI (D2G-SOI) process functions to 0.7V in [116]. The authors analyze their design for bulk CMOS and report that it cannot operate below 1.0V [116]. A bulk CMOS 65nm SRAM also reports its minimum operating voltage as 0.7V [107].

Our results confirm that SNM degradation and inability to write are the two most significant obstacles to sub-threshold SRAM functionality in 65nm. We examine the SNM problem in greater detail in Section 5.3. This section examines the critical problems with write and read operation for a 65nm 6T SRAM in sub-threshold.

5.2.1 Write Operation

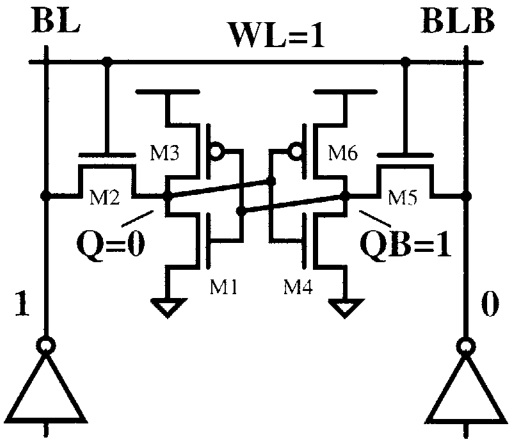


Figure 5-15: Schematic showing conceptual write of '1' into node 'Q' in the 6T bitcell. Data shown for Q and QB must be overwritten.

Figure 5-15 shows a 6T bitcell at the onset of a write operation. The write drivers are applying the new bit values to the bitlines, but the cell still holds the old values. In order for Q and QB to convert from their old values ('0' and '1') to their new values ('1' and '0'), the write drivers must overpower the feedback inside the cell. Since the nMOS access transistors, M_2 and M_5, are poor at driving a '1', the key to proper write operation lies with writing the new '0' correctly. This requirement presents a well-known sizing problem for above-threshold design. Specifically, there is a ratioed fight between the pMOS inside the cell that holds a '1' and the series combination of the nMOS access transistor and the write driver [98]. In Figure 5-15, the write driver on BLB and M_5 must overpower M_6 to write a '0' to node QB.

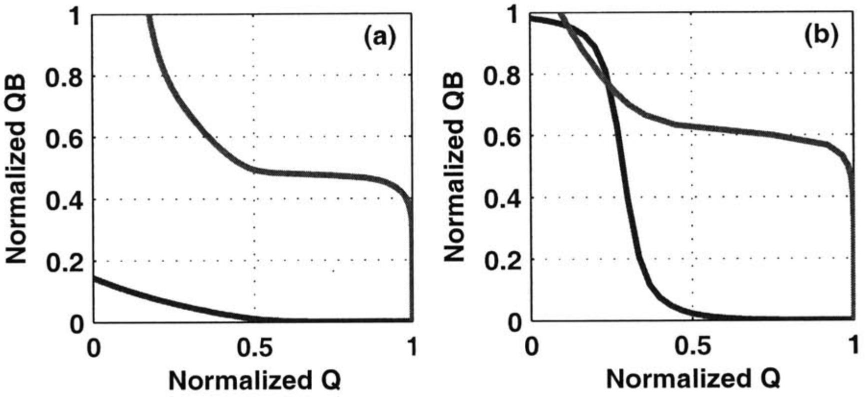


Figure 5-16: Example VTCs for write access showing write 'SNM'. A negative write SNM indicates a successful write to the mono-stable point (a). If the write SNM remains positive, then the write fails because the state of the bitcell is not changed - example shown in (b).

When the write operation succeeds, the bitcell becomes mono-stable, forcing the internal voltages to the correct values. Figure 5-16(a) shows a butterfly curve of a bitcell displaying correct write operation. This is the same type of plot used to demonstrate SNM for the bitcell. During a successful write, there are no lobes on the butterfly curve. Using the SNM terminology, we can say that the 'write SNM' is negative. If the VTC and inverse VTC curves on the plot shift by an amount equal to the write SNM, then the cell will regain bi-stability. Figure 5-16(b) gives a different example of a positive write SNM that corresponds to a failure to write the bitcell. In this particular example, the static characteristics of the butterfly curve show that the write driver and access transistor cannot sufficiently overpower the bitcell's feedback to write Q='1'. The key to achieving a successful write in the traditional fashion, then, is to ensure that the access transistor and write driver win the fight with the pMOS pull-up inside the bitcell.

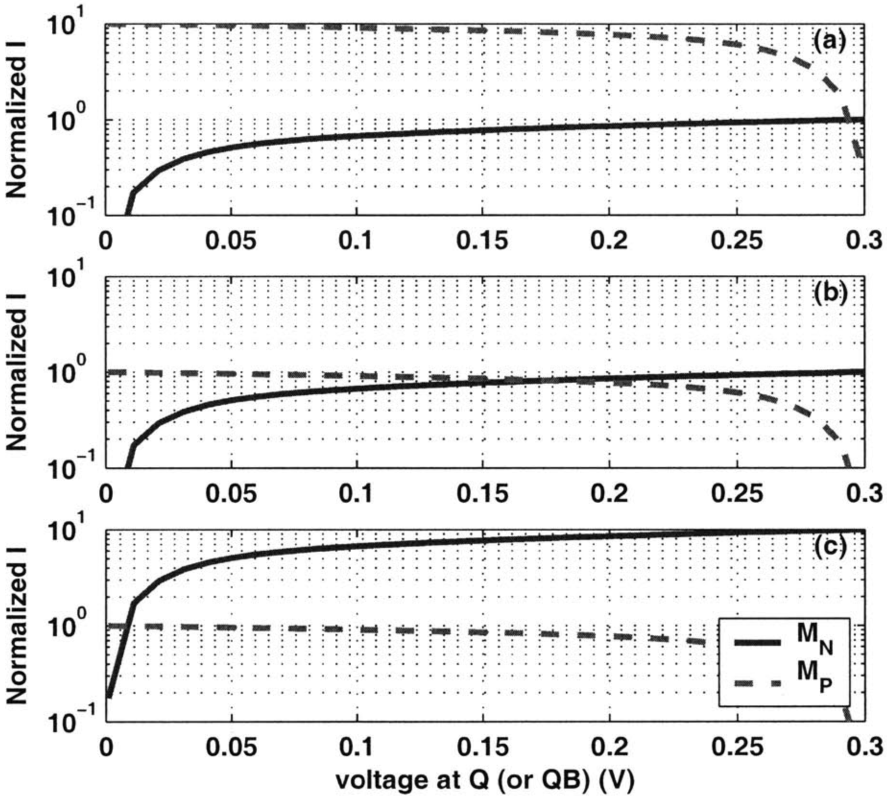


Figure 5-17: Plot showing the impact of relative pMOS and nMOS on-current on voltage at internal bitcell node. $I_P > I_N$ (a), $I_P \approx I_N$ (b), $I_P < I_N$ (c).

For above-threshold operation, nMOS current usually exceeds pMOS current by a few times for iso-size because of the difference in device mobility. The sizing difference between M_6 and M_5 (or M_3 and M_2) therefore does not need to be large to ensure that the access transistor can pull down the internal node during a write. In sub-threshold operation, however, this is a different matter. The ratio of currents in a pMOS and nMOS of the same size depends exponentially on the threshold voltage,

V_T. Figure 5-17 conceptually shows the impact of different relative strengths in the transistors. It shows the load line analysis for M_6 and M_5 during a write access, assuming that both transistors have $|V_{GS}| = V_{DD}$. If, as in Figure 5-17(a), the pMOS current is significantly higher, then node QB will not pull to '0', and we cannot write the cell. In Figure 5-17(b) and (c), where the nMOS is strong relative to the pMOS, we can ensure that the internal node gets pulled to '0' for a successful write operation.

The main problem in sub-threshold is that any of these scenarios shown in Figure 5-17 can occur even in the same process. Even if the pMOS and nMOS currents are well-balanced at the TT corner, process variation can easily create a relative difference in pMOS and nMOS current of an order of magnitude or more. Furthermore, local variations in the V_T of transistors from cell to cell can aggravate this problem. As we described in an earlier chapter, sizing alone is not a strong knob for fixing this problem because only unreasonable sizing ratios could account for the wide ranges of possible current that arise due to V_T mismatch.

In the 65nm process for which we are designing, the pMOS mobility makes it weaker than an iso-sized nMOS at nominal V_{DD}, but the pMOS relationship to the nMOS in sub-threshold is more nearly that of Figure 5-17(a). This makes write functionality more challenging. Figure 5-18 shows the write margin of a 6T bitcell versus temperature and process corner. Again, negative SNM corresponds to correct write functionality. At V_{DD} = 300mV in Figure 5-18(a), the standard method of writing fails for large regions of process corner and temperature. The general trend showing an improvement of write operation (i.e. more negative write margin) at higher temperature occurs because the pMOS transistors weaken relative to nMOS as temperature rises. Thus, the access transistors become more capable of overcoming the pMOS that holds a '1' inside the bitcell. As supply voltage increases, the write margin improves. Figure 5-18(b) shows the write margin at 0.6V. This voltage is well above the V_T of both types of transistor, and the pMOS has weakened relative to the nMOS because the mobility starts to dominate the differences in V_T. Even at 0.6V, the write margin is barely negative for the worst-case corner, and this plot does not account for local V_T variation. For these reasons, V_{DD}=0.6V is the best case voltage

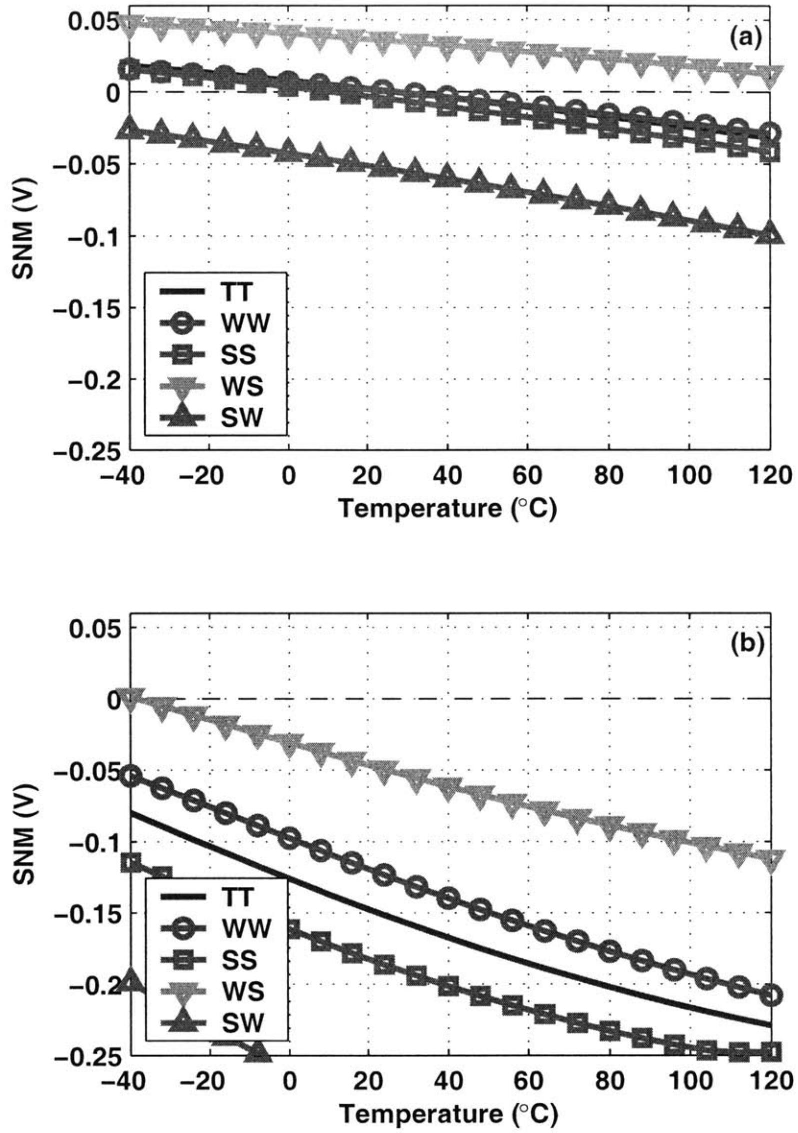


Figure 5-18: SNM for write access versus temperature and process corner (TT, WW, SS, WS, and SW) at $V_{DD} = 0.3\text{V}$ (a) and $V_{DD} = 0.6\text{V}$ (b). Negative SNM indicates successful write.

for which we can expect traditional write operations to work for a sub-threshold memory in this 65nm process.

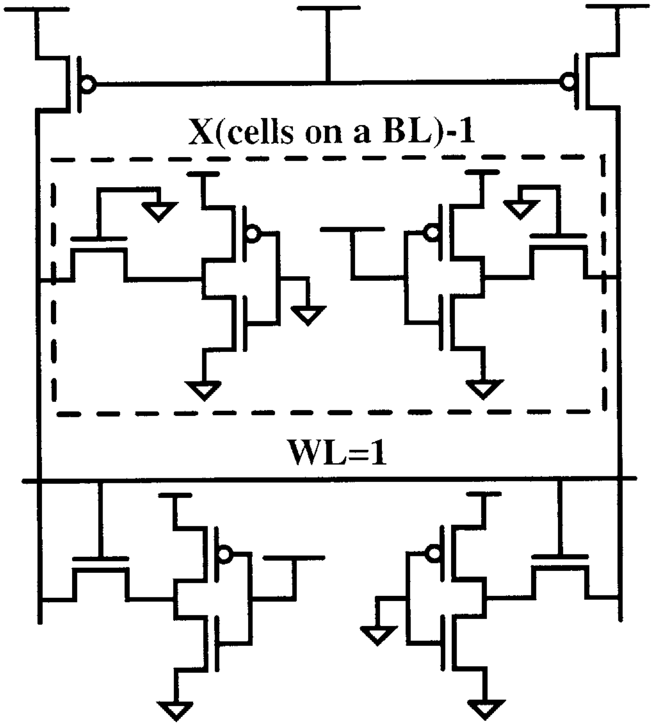


Figure 5-19: Schematic showing the simulation set-up for finding steady-state behavior for the worst-case scenario for bitline leakage. For X cells on a bitline, $X - 1$ cells hold the complement of the value in the accessed cell, maximizing leakage in opposition of the desired read.

5.2.2 Read Operation

As described in Section 5.1.4, bitline leakage is a significant problem for DSM SRAM. Despite various techniques to minimize leakage into unaccessed cells, technology scaling leads to progressively shorter bitlines that have been segmented to reduce bitline leakage [97]. This section examines the impact of bitline leakage on read operation in the sub-threshold region for a 65nm technology.

Figure 5-19 shows the simulation set-up for examining the steady-state behavior for the worst-case bitline leakage. After the read access reaches steady state, the on-current of the access transistors for the addressed bitcell drives the bitlines. The leakage current of the remaining $X - 1$ bitcells opposes the accessed cell. The dc voltage on the bitlines shows the steady-state impact of bitline leakage for X cells on a bitline.

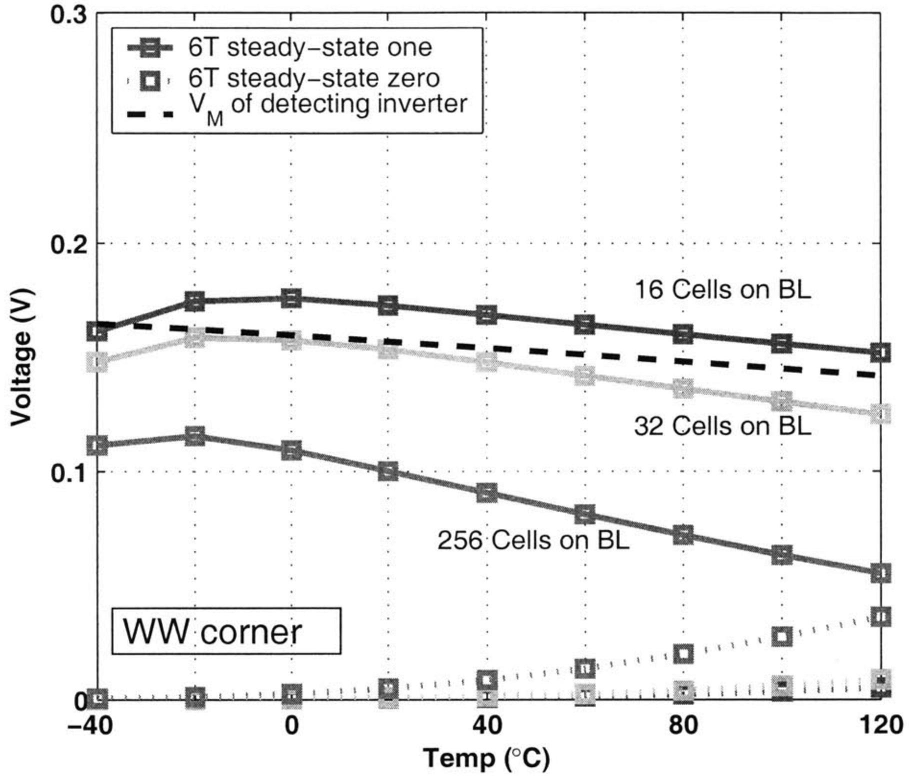


Figure 5-20: Simulation of scenario in Figure 5-19 showing steady-state bitline voltages. Bitline leakage severely limits the number of cells that can share a bitline.

Figure 5-20 plots the results of a simulation using the set-up from Figure 5-19 for the 6T bitcell in 65nm CMOS at V_{DD} = 300mV. The bitline that should be '1' droops significantly because of bitline leakage into the other cells. This figure shows the WW corner which shows the worst-case, but the other corners do not give significant improvement. The droop gets worse at higher temperature because the pMOS transistors weaken relative to nMOS as temperature increases, strengthening the bitline leakage. The plot shows the switching threshold, V_M, of an inverter that can serve as a simple sense amplifier to detect the full-swing output of this bitcell. Clearly, a 6T SRAM using this inverter for sensing is limited to 16 bitcells on a bitline at best. Even more complicated sense amplifiers face the challenge of operating on a small bitline differential that develops slowly because of leakage.

The other, more fundamental problem for 6T bitcells in sub-threshold is degraded Read SNM. The next section deals with SNM in greater detail.

5.3 Static Noise Margin

This section evaluates the Static Noise Margin (SNM) of 6T SRAM bitcells operating in sub-threshold. We analyze the dependence of SNM during both hold and read modes on supply voltage, temperature, transistor sizes, local transistor mismatch due to random doping variation, and global process variation in a commercial 65nm technology. We analyze the statistical distribution of SNM with process variation and provide a model for the tail of the Probability Density Function (PDF) that dominates SNM failures.

The minimum voltage for retaining bistability was theorized in [46] and modeled for SRAM in [96], but degraded SNM can limit voltage scaling for SRAM designs above this minimum voltage. SNM quantifies the amount of voltage noise required at the internal nodes of a bitcell to flip the cell's contents.

An expression for above-threshold SNM based on long-channel models is given in [6], and [139] models above-threshold SNM for modern processes with process variation. This section builds on previous work by examining SNM for sub-threshold SRAM [140].

5.3.1 Modeling Sub-threshold Static Noise Margin

As we discussed in Chapter 2, lowering V_{DD} reduces gate current much more rapidly than sub-threshold current, so total current in the sub-threshold region can be modeled to first order as in (5.1).

$$I_D = I_S \exp\left(\frac{V_{GS} - V_T}{nV_{th}}\right)\left(1 - \exp\left(\frac{-V_{DS}}{V_{th}}\right)\right) \tag{5.1}$$

The sub-threshold factor $n = 1 + C_{ds}/C_{ox}$, $V_{th} = kT/q$, and I_S is the current when V_{GS} equals V_T. For simplicity, we treat pMOS parameters as positive values.

For the 65nm technology used in this section, the nMOS drive current is higher in above-threshold than the pMOS for iso-width, but the pMOS current is higher in sub-threshold. During hold mode, $WL = 0$ so M_2 and M_5 (refer to Figure 5-1) have $V_{GS} \leq 0$ and thus negligible current. We can model the cell VTCs ($V_{OUT} = f_{VTC}(V_{IN})$) as those of a simple inverter in sub-threshold.

$$\begin{aligned} V_{QB} =& V_{th}\frac{n_1 n_3}{n_1+n_3}\left(\ln\frac{I_{S3}}{I_{S1}} + \ln\left(\frac{1-\exp((-V_{DD}+V_Q)/V_{th})}{1-\exp(-V_Q/V_{th})}\right)\right) \\ &+ \frac{n_1 V_{DD}}{n_1+n_3} + \frac{n_1 n_3}{n_1+n_3}\left(\frac{V_{T1}}{n_1} - \frac{V_{T3}}{n_3}\right) \end{aligned} \tag{5.2}$$

Referring to Figure 5-4, equation (5.2) [46] gives the inverse VTC for inverter 1 ($V_{IN} = f_{VTC}^{-1}(V_{OUT})$). The inverse of (5.2) is given in [67] for matched pMOS and nMOS (same n, V_T,I_S). We give a full solution for $V_{OUT} = f_{VTC}(V_{IN})$ for inverter 2 in (5.3).

$$V_{QB} = V_{DD} + V_{th}\ln\left(\frac{1-G+\sqrt{(G-1)^2+4\exp(\frac{-V_{DD}}{V_{th}})G}}{2}\right) \tag{5.3}$$

$$G = \exp\left(\frac{n_4+n_6}{n_4 n_6 V_{th}}Q - \ln\frac{I_{S6}}{I_{S4}} - \frac{V_{DD}}{n_6 V_{th}} - \frac{1}{V_{th}}\left(\frac{V_{T4}}{n_4} - \frac{V_{T6}}{n_6}\right)\right) \tag{5.4}$$

Figure 5-21(a) plots (5.2) and (5.3) against simulation curves for no local mismatch and for 1σ V_T mismatch in M_6. The steeper slope of the VTC transition for the model results from not modeling the DIBL effect.

During a read access, $WL = V_{DD}$ and the bitlines are precharged to V_{DD} so, if $V_Q = 0$ prior to access, M_1 and M_2 are both on. This creates a voltage division that raises the voltage at Q. Assuming pMOS current is negligible in the region of interest, (5.5) shows the inverse VTC equation during a read operation near the SNM [8] for inverter 1.

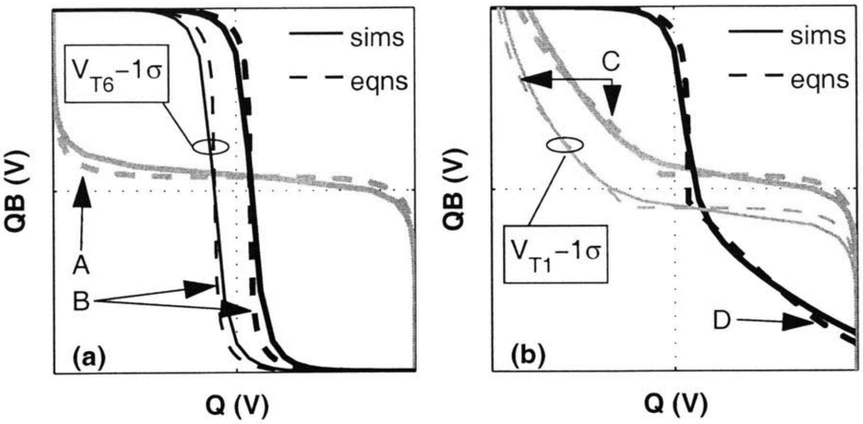


Figure 5-21: First-order VTC equations versus simulation for hold (a) and for read (b). Line A is (5.2), line B is (5.3), line C is a piecewise combination of (5.5) and (5.2), and line D is a piecewise combination of (5.3) and the graphical inverse of (5.5).

$$\begin{aligned} V_{QB} =& n_1 V_{th} \ln \frac{I_{S2}}{I_{S1}} + n_1 V_{th} \ln \left(\frac{1 - \exp((-V_{DD} + V_Q)/V_{th})}{1 - \exp(-V_Q/V_{th})} \right) \\ &+ V_{T1} + \frac{n_1}{n_2} (V_{DD} - V_{T2} - V_Q) \end{aligned} \tag{5.5}$$

This equation cannot be inverted analytically, and it applies only to the region of the VTC where V_{OUT} is low. Figure 5-21(b) shows (5.5) and its graphical inverse combined piecewise with (5.2) and (5.3) and plotted against simulation for no local mismatch and for 1σ V_T mismatch in M_1 for minimum device sizes at 25^oC.

Graphical or numerical solutions for SNM are easily derived from the VTC equations, although no direct analytical solution exists. The equations provide a good estimate of the behavior of the SNM based on key parameters. One shortcoming of (5.2)-(5.5) is the assumption that sub-threshold slope ($S = nV_{th} \ln 10$) is constant for each transistor. Figure 5-22(a) shows that S varies with V_{GS}, and Figure 5-22(b) shows S changing with temperature without the expected constant slope due to V_{th}. A more crucial problem with (5.2)-(5.5) is the assumption that certain currents are negligible. These assumptions break down under certain combinations of V_T variation, rendering the first-order equations inaccurate.

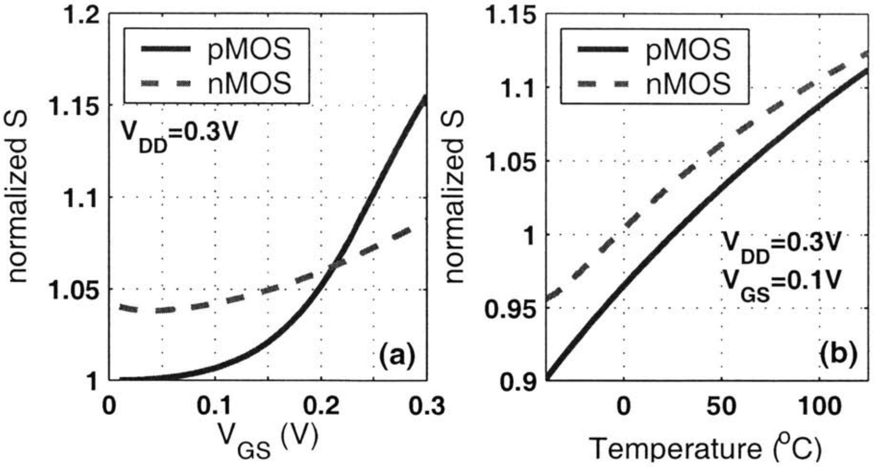


Figure 5-22: Changes in sub-threshold slope (S) versus V_{GS} (a) and temperature (b).

5.3.2 Sub-threshold SNM Dependencies

With embedded SRAM often providing multiple megabits of storage, the SNM of the nominal bitcell becomes largely irrelevant. Variations in processing and in the chip's environment create a distribution of SNM across the bitcells in a given memory, and the worst-case tail of this distribution determines the yield. This section examines the impact of different parameters on SNM in sub-threshold and offers a model for estimating the tail of the SNM density function for process variation.

SNM for a bitcell with ideal VTCs is still limited to $V_{DD}/2$ because of the two sides of the butterfly curve. An upper limit on the change in SNM with V_{DD} is thus $\frac{1}{2}$. Figure 5-23 shows example butterfly curves at different supply voltages from 1.2V to 200mV for both hold (a) and read (b). Figure 5-24 plots SNM versus V_{DD} directly for both hold and read mode. The slopes of the curves confirm that less than $\frac{1}{2}$ of V_{DD} noise will translate into SNM changes.

The impact of temperature on SNM in sub-threshold is not too significant because the ratio of nMOS and pMOS current does not change widely. Figure 5-25 shows SNM versus temperature in sub-threshold and again for strong inversion. The sensitivity in sub-threshold is lower, and varying temperature from -40^oC to 125^oC only alters Read and Hold SNM by 21mV and 6mV, respectively. Higher temperatures lower

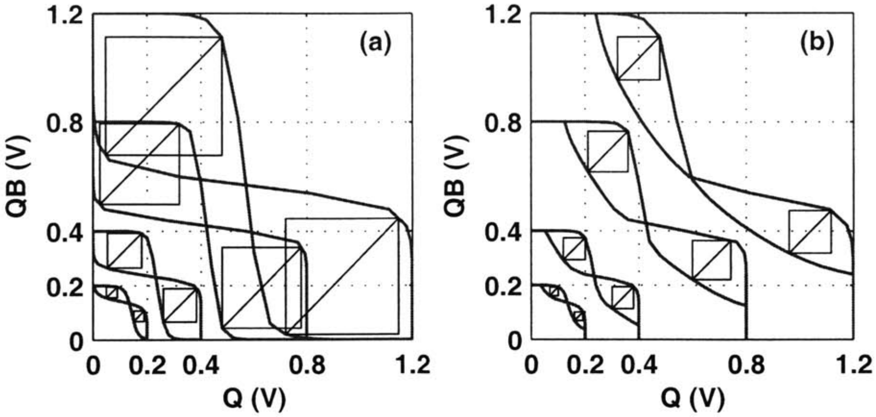


Figure 5-23: VTCs for Hold (a) and Read (b) with varying V_{DD}.

SNM in sub-threshold due to the degraded gain in the inverters that results from worse sub-threshold slope (see Figure 5-22(b)). Also, pMOS devices weaken relative to nMOS at higher temperature. Figure 5-26 provides example butterfly plots for 0^oC and 100^oC.

In contrast to above-threshold [141], Figure 5-27 shows that cell ratio $((\frac{W}{L})_1/(\frac{W}{L})_2$ or $(\frac{W}{L})_4/(\frac{W}{L})_5)$ has very little impact on SNM during sub-threshold read. In fact, sub-threshold SNM sensitivity to any sizing changes is reduced. The lower impact

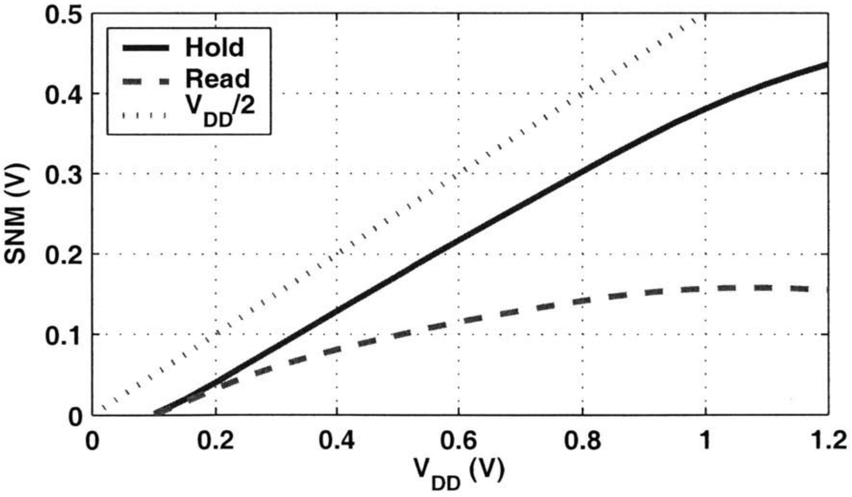


Figure 5-24: SNM versus V_{DD}

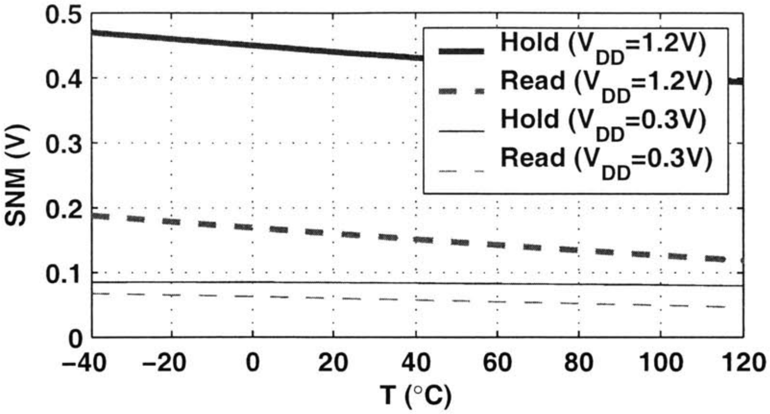


Figure 5-25: SNM versus temperature.

of sizing is intuitively reasonable considering the exponential dependence of subthreshold current on other parameters. Mathematically, we can see from (5.2)-(5.5) that sizing changes affect I_{Si} linearly and only have a logarithmic impact on the VTCs. One point of caution here is that V_T for deep submicron devices tends to vary with size as a result of narrow or short channel effects. The impact of this V_T change

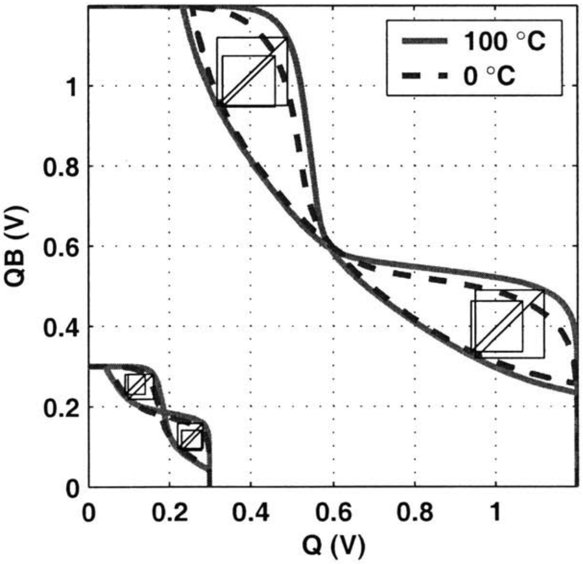


Figure 5-26: VTCs during a read access across temperature.

that might accompany a sizing change is more pronounced. These effects depend on the technology and make general SNM modeling more complicated.

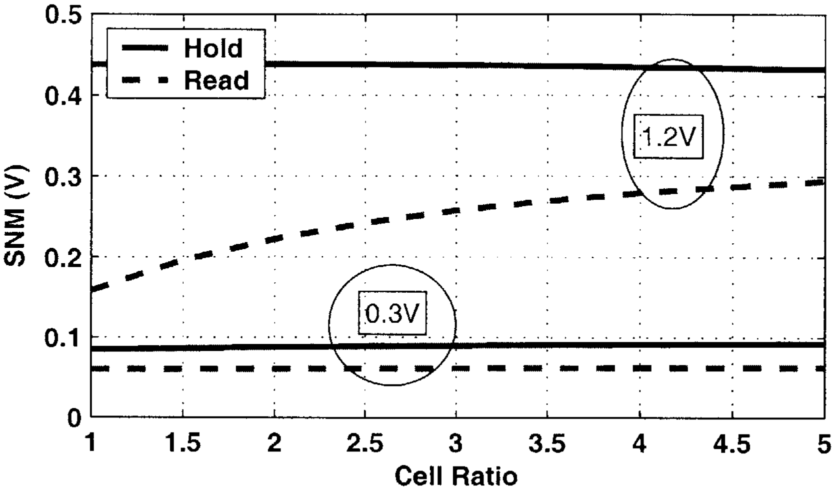


Figure 5-27: Cell ratio affects SNM less in sub-threshold

5.3.3 Dependence on Random Doping Variation

The randomness of the number of doping atoms and their placement in a MOSFET channel causes random mismatch even in transistors with identical layout [142]. The impact on threshold voltage, whose σ is proportional to $(WL)^{-\frac{1}{2}}$, is the worst for minimum sized devices which are common in SRAM. Local variation is a huge problem for SRAM functionality, and it is the subject of many papers (e.g. [143][116]). The exponential dependence of current on V_T in sub-threshold operation makes this random variation even more influential. Furthermore, the large number of bitcells in many SRAMs makes the tails ($5-6\sigma$) of the PDF more critical for modeling since the extreme cases are the limiting factor for yield. Previous work has shown that above-threshold SNM is nearly linear with V_T, and modeling its slope as constant allows an approximation of the joint PDF for SNM [139]. Likewise, the sensitivity of above-threshold SNM to V_T is linearized for each transistor in [144].

Figure 5-28 shows that, like in strong inversion, the sensitivity of SNM high (the

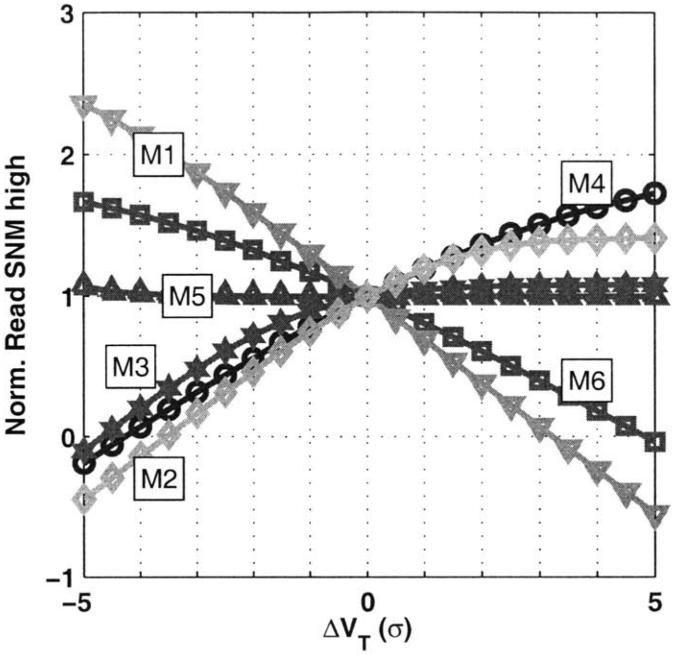


Figure 5-28: Dependence of SNM high on single FETs is nearly linear.

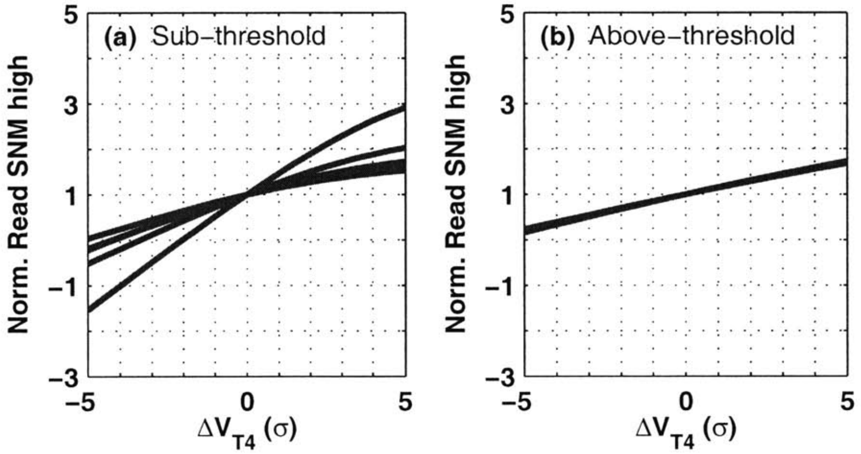


Figure 5-29: Dependence of SNM high on a single FET depends on other V_Ts in sub-threshold (a), unlike for above-threshold (b).

upper-left box in Figure 5-7) is nearly linear with each individual V_T. However, Figure 5-29(a) shows the relationship between SNM and V_{T4} for a few different values of the other V_Ts. The obvious dependence of the slope on the other V_Ts prevents using a model of the form $SNM = SNM_0 + \sum c_i V_{Ti}$ for sub-threshold SNM. The same is not true of above-threshold, shown in Figure 5-29(b), for which a first order series model works well [144][139].

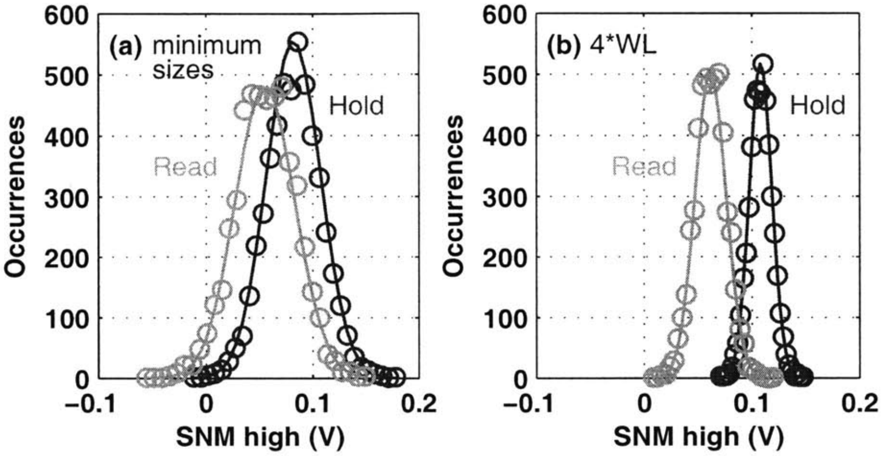


Figure 5-30: SNM high and low (not shown) for a minimum sized cell (a) and for 4*W*L (b) is normally distributed with random V_T mismatch in all transistors.

Figure 5-30 shows the results of 5k-point Monte-Carlo (M-C) simulations with random independent V_T mismatch in all transistors. These histograms confirm that sub-threshold SNM at the upper lobe of the butterfly curve (SNM high) is normally distributed. The solid lines show a fitted Gaussian PDF, and the markers show simulation results. Larger sizes for the bitcell clearly have the advertised effect of lowering the variance of V_T as seen in Figure 5-30(b). The SNM low PDFs are essentially identically distributed. The scatter plot in Figure 5-31 shows that SNM high and SNM low are correlated. The dependencies for mismatch in each single transistor are overlaid in white for reference. The Hold SNM shows a saturation effect along the upper edge. SNM high and SNM low are not independent because any change to a VTC that increases the SNM at one side tends to decrease SNM at the other side.

The actual SNM that matters for a bitcell is the minimum of SNM high and SNM low. Thus, the random variable $X_{SNM} = \min(X_{SNMhigh}, X_{SNMlow})$. Order statistics can provide us with the PDF for the minimum of n independent, identically

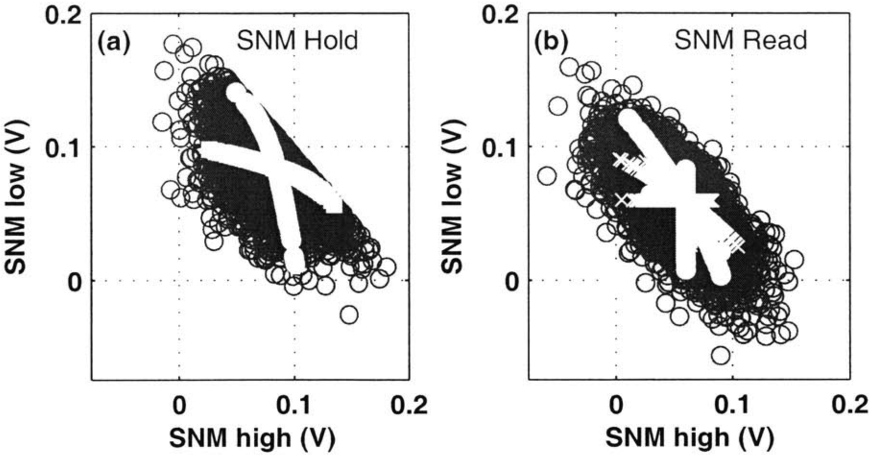


Figure 5-31: Scatter plots for SNM high vs. SNM low with single FET dependencies overlaid in white.

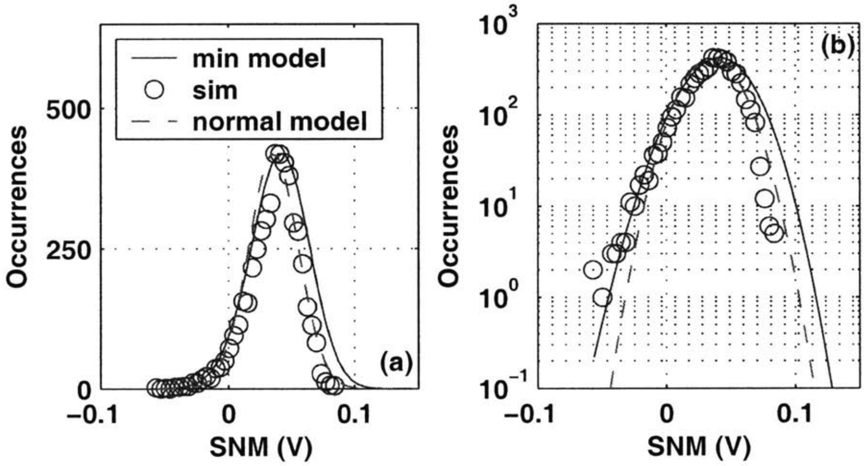


Figure 5-32: Histogram of Read SNM Monte-Carlo simulation (circles) with normal PDF (dash) and PDF based on (5.7) (solid) over-laid. The semilog plot (b) shows that the PDF based on (5.7) matches the worst-case tail quite well.

distributed (*iid*) random variables, X_i. If f is the PDF, and F is the Cumulative Distribution Function (CDF) for X_i, the PDF of the minimum of two *iid* variables is given in (5.6).

$$f(\min(X_1, X_2)) = 2f_X(1 - F_X) \tag{5.6}$$

Although SNM high and SNM low are normally distributed with approximately the same mean and variance, we have previously shown that they are not independent. However, we are less interested in modeling the entire PDF for SNM than we are in modeling the worst-case tail. As previously stated, the tail toward lower SNM is the limiting factor. Let us assume that they are *iid*. Then we can solve for the PDF as:

$$f_{SNM} = 2f_{SNMhigh}(1 - F_{SNMhigh}) \tag{5.7}$$

and the CDF is simply:

$$F_{SNM} = 2F_{SNMhigh} - (F_{SNMhigh})^2 \tag{5.8}$$

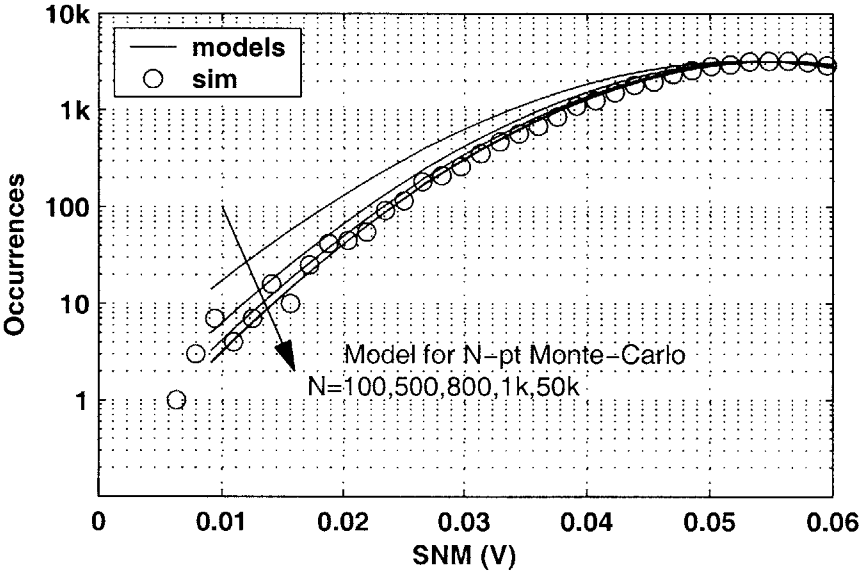


Figure 5-33: 50k-point Monte-Carlo simulation for SNM with 4*WL sized transistors. Model based on 1k-point Monte-Carlo data matches the 50k-point model with < 3% error.

Figure 5-32 shows the histogram for a 5k-point M-C simulation of Read SNM plotted on linear axes (a) and semilog axes (b). Clearly, SNM is not normally distributed, and its mean is lower than the mean of SNM high and SNM low. Figure 5-32(b) shows that a Gaussian PDF does not match the worst-case tail on the left side of the PDF. On the other hand, the PDF based on (5.7) provides a good estimate of the worst-case tail. The plot shows that the model does not fit the distribution above the mean. This shortcoming results from the correlation between SNM high and SNM low. Since these two random variables are not *iid*, we cannot claim that the minimum model will always match the tail. However, we can show experimentally that it does offer a good estimate. Thus, the model is a useful tool for evaluating SNM under different design decisions and conditions. This PDF gives the powerful option of estimating the SNM at the worst-case end of the PDF without using extremely long M-C simulations until the design space is narrowed sufficiently.

Figure 5-33 shows several estimated PDFs using (5.7) that are based on data sets of different lengths. These estimates are plotted over a 50k-point M-C simulation. A 1000-point M-C simulation gives an estimate that overlays the estimate from the 50k-point case on the plot ($<$ 3% error). Using this approach allows a designer to reliably estimate the tail of the SNM PDF for a large memory with relatively few samples.

Thus far we have assumed that device mismatch occurs in transistors that start off as typical for the process. In addition to the inter-die V_T mismatch that we have described is an intra-die process variation that sets the process corner (e.g. fast nMOS, slow pMOS, etc.). Even with no local transistor variation, the global process corner impacts the SNM. Figure 5-34 shows the SNM PDF for a minimum sized 6T bitcell from a M-C simulation of global process corner in which nine process parameters are varied. Here again, the tail of the PDF is the limiting factor.

In a production framework, each die containing a given SRAM will have a global process corner that affects SNM as in Figure 5-34. On top of this, mismatch in each cell will result from random doping variation. Assuming that any die within 3σ of the mean is usable, we found the global process corner that gives an SNM yield with the same probability as -3σ for both hold and read cases. Figure 5-35 shows that the impact of mismatch at this 3σ process corner is essentially to shift the mean of the PDF by the offset caused by global variation. This means that the models we have presented remain valid for the case of combined global and local variation. Figure

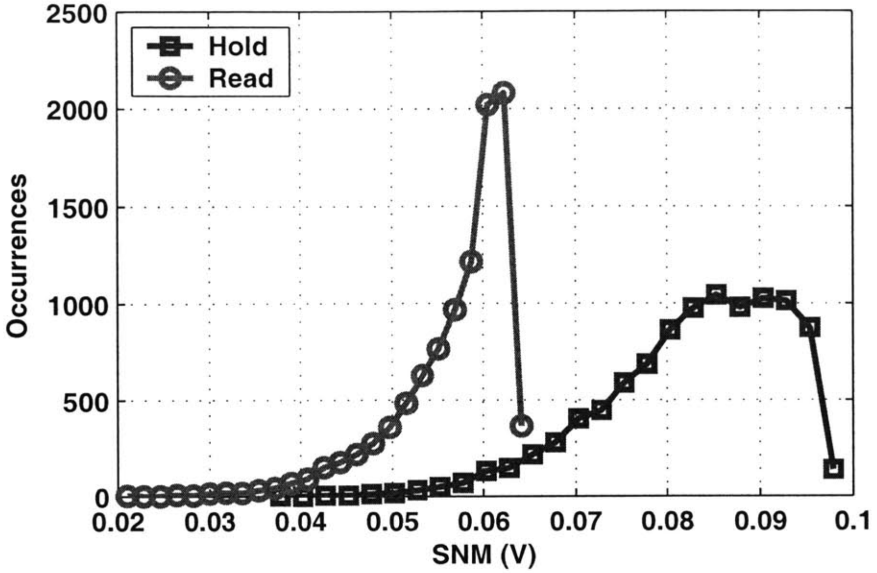


Figure 5-34: Monte-Carlo simulation showing global variation impact on SNM for a minimum sized bitcell.

5-36 shows the semilog plot of the distributions to confirm this conclusion.

Static noise margin is a critical metric for SRAM bitcell stability. This section has explored the impact of different parameters on SNM for SRAM bitcells in sub-threshold. The dominant factor affecting sub-threshold circuits in general and SNM specifically is V_T mismatch due to random doping variation, and the critical region for examination is the tail of the SNM PDF. We have shown that first-order theoretical models for calculating SNM are accurate close to the nominal values of V_T, but they cannot accurately account for all of the mismatch cases. We have shown that SNM high and SNM low are normally distributed with local V_T variation and correlated. Despite their correlation, we have shown that treating them as *iid* leads to a PDF for SNM that gives an accurate model of the tail cases. This estimate is invaluable for avoiding long Monte-Carlo simulations in the design of large SRAMs for sub-threshold operation.

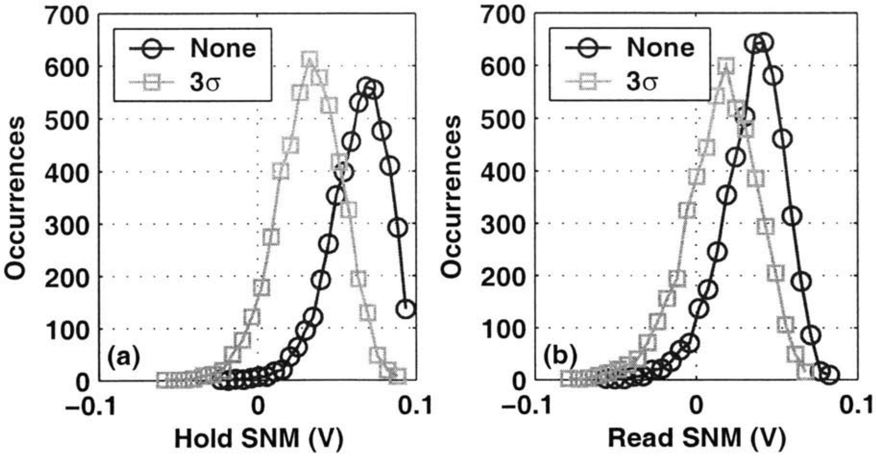


Figure 5-35: SNM Monte-Carlo simulations for local mismatch on top of global variation.

5.3.4 Implications for Sub-threshold SRAM

As described in the preceeding section, the impact of local V_T variation on Read SNM imposes a serious limitation on SRAM V_{DD} scaling. Increasing the sizes of the transistors in the bitcell and/or raising V_{DD} is necessary to achieve the desired statistical yield. This problem suggests that changing the bitcell to eliminate the Read SNM problem would allow lower V_{DD} operation and thus decrease both total energy and standby power.

Figure 5-37 shows the distribution of the Read and Hold SNMs for a 6T bitcell at a 300mV supply voltage. The mean Read SNM is only slightly above half of the mean Hold SNM, but, even worse, the deviation of the Read SNM is larger than for the Hold SNM. For a multiple megabit memory, numerous cells will have Read SNM less than zero based on this statistical analysis. From this figure, the mean of the Read SNM at 500mV roughly equals the mean of the Hold SNM at 300mV. However, it is unclear from this plot how the Hold SNM and Read SNM compare at the worst-case tails.

Figure 5-38 shows the CDF functions derived from the distributions using (5.8). These CDF curves show how Hold SNM compares to Read SNM. For 6σ probability,

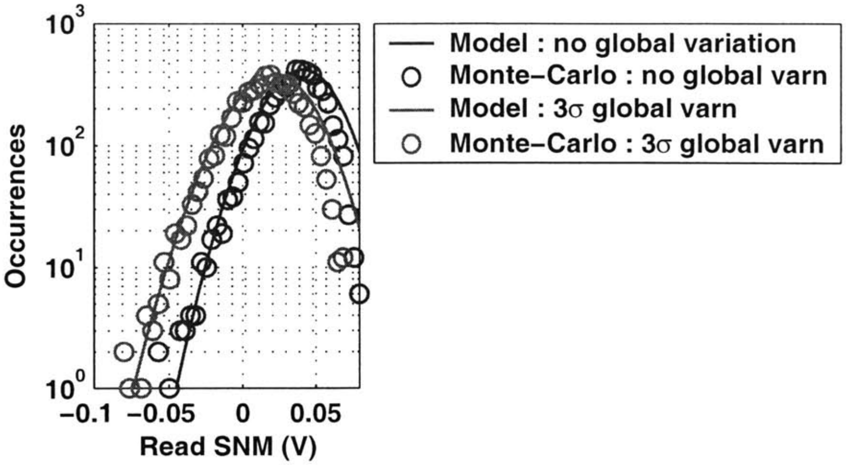


Figure 5-36: SNM Monte-Carlo simulations for local mismatch on top of global variation (none or 3σ) compared to the model for Hold (a) and Read (b).

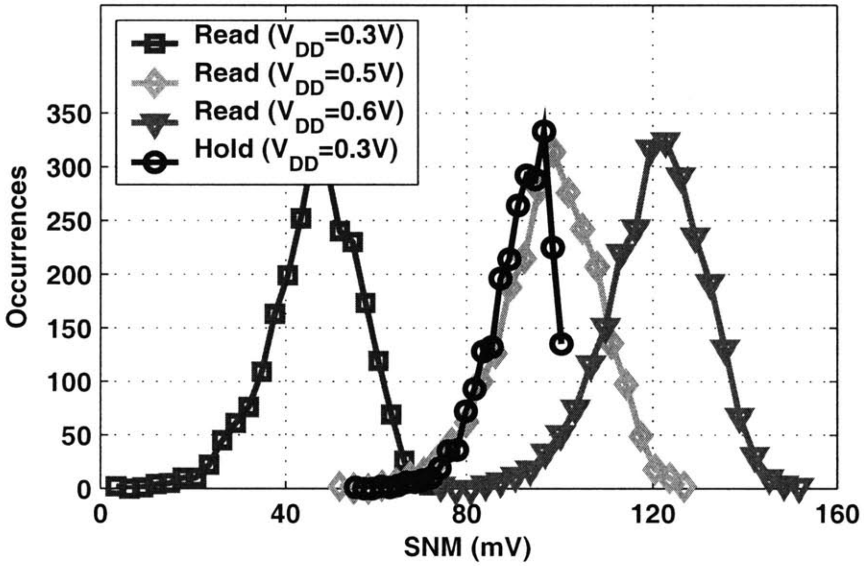


Figure 5-37: Distribution of Hold SNM at 300mV compared with Read SNM distributions at different voltages. Read SNM at 500mV has the same mean, but it has a larger standard deviation.

the Hold SNM for a given V_{DD} roughly equals the Read SNM for twice that V_{DD} in the range of interest for us. This means that a memory that avoids the Read SNM problem can operate at roughly half of the V_{DD} of a 6T memory with the same 6σ

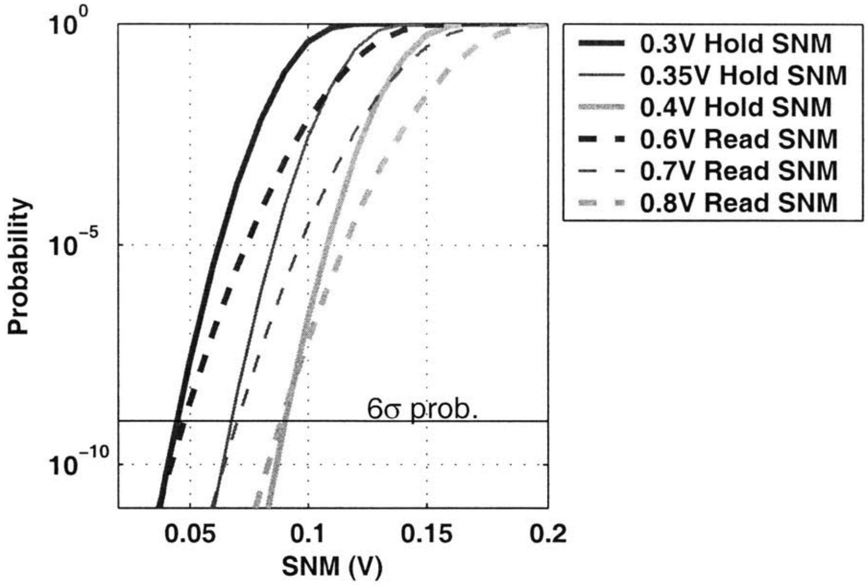


Figure 5-38: CDFs of SNM distributions showing that avoiding the Read SNM allows a reduction in V_{DD} by ~ 0.5 for the same 6σ stability.

bitcell stability. The fact that by avoiding the limitation imposed by Read SNM we can operate at lower voltages with the same cell stability is a key observation used in the design of a sub-threshold bitcell.

5.4 A Sub-threshold Bit-cell Design

The previous sections in this chapter point to several advantages of an SRAM bitcell that can operate into the sub-threshold region. Previously published works have scaled SRAM V_{DD} into the sub-threshold region during idle, but no SRAM actually operates in this region. The 0.18μm memory in [4] provides one exception. This memory operates down to 180mV, deep into the sub-threshold region. In structure, it resembles a register file more nearly than a standard 6T SRAM. The bitcell itself looks like a latch with a tristate driver for writing into the cell and a tristate inverter replacing a standard inverter in the cross-coupled inverter pair so that it can cut off the feedback during write. For read, the outputs of a pair of cells are multiplexed together, and these outputs are successively multiplexed until the addressed word is selected. These muxes correspond to a bitline shared by only two bitcells. Accounting for the multiplexors required to read, the equivalent bitcell size is 18 transistors. Despite its large size, this sub-threshold SRAM benefits from the robustness of full-swing read and static, non-ratioed write operations.

Taking this previous implementation [4] as a datapoint in the set of possible bitcells, we can set up a range of bitcell options. At one end of the range, the 6T bitcell cannot operate below 600-700mV in 65nm. At the other end of the spectrum, an 18T bitcell will function robustly in sub-threshold since it looks and functions very much like combinational logic. Along the range in between these two options are many possible bitcell designs that address the obstacles to sub-threshold operation by increasing the number of transistors relative to the 6T cell. The bitcell that this section describes was selected from among many others because it represents the best trade-off of functionality and area. In other words, it is the smallest bitcell from those examined that provides robust sub-threshold functionality.

Before describing the sub-threshold bitcell, we will briefly review the key problems facing the 6T cell for sub-threshold operation. First, the mismatch of current in sub-threshold prevents the standard write operation from overpowering the cell's feedback. In this technology, the standard write approach cannot function robustly below 600mV (see Section 5.2.1). Secondly, the Read SNM depends strongly on local variation in the bitcell. Large memories cannot hope to function below 600-700mV without a significant number of statistical failures during read access (see Section 5.3). Finally, even if Read SNM was not a problem, bitline leakage severely hampers the read operation and limits the number of cells shared on a bitline to at most 16 (see Section 5.2.2).

Figure 5-39 shows the schematic of a 10T bitcell that addresses these problems and provides sub-threshold functionality. Transistors M_1 through M_6 are identical to a 6T bitcell except that the source of M_3 and M_6 tie to a virtual supply voltage rail, VV_{DD}. Write access to the bitcell occurs through the write access transistors, M_2 and M_5, from the write bitlines, BL and BLB. Transistors M_7 through M_{10} implement a buffer

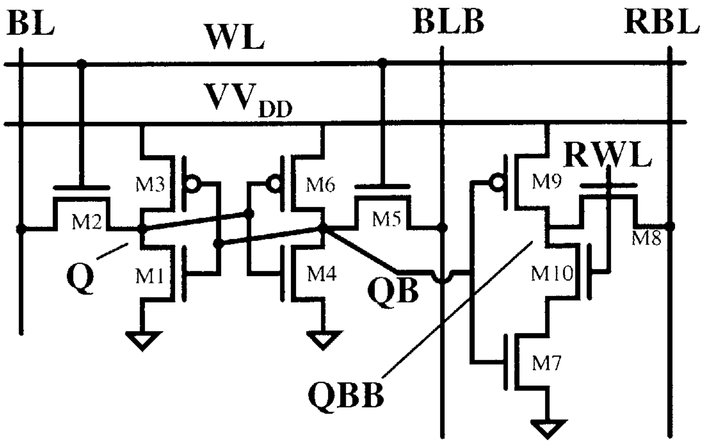


Figure 5-39: Schematic of the 10T sub-threshold bitcell.

used for reading. Read access is single-ended and occurs on a separate bitline, RBL, which is precharged prior to read access. The wordline for read also is distinct from the write wordline. One key advantage to separating the read and write wordlines and bitlines is that a memory using this bitcell can have distinct read and write ports. Since a 6T bitcell does not have this feature, the 10T bitcell is in some ways more fairly compared to an 8T dual-port bitcell (6T bitcell with two pairs of access transistors and bitlines). The remainder of this section describes the operation of this bitcell in detail.

5.4.1 Enabling Sub-threshold Read

The 10T bitcell in Figure 5-39 uses transistors M_7-M_{10} to remove the problem of Read SNM by buffering the stored data during a read access. When the read wordline (RWL) goes high, the pre-charged read bitline (RBL) causes a voltage divider across M_7, M_8, and M_{10}, but this increase in voltage at QBB does not impact the stored data at Q and QB. Thus, the worst-case SNM for this bitcell is the Hold SNM related to M_1-M_6, which is the same as the 6T Hold SNM for same-sized M_1-M_6. As described in Section 5.3.4, eliminating the Read SNM problem allows this bitcell to operate at half of the V_{DD} of a 6T cell while retaining the same 6σ stability. A different approach for eliminating the Read SNM in [145] uses a 7T cell to prevent the higher voltage at the internal node from propagating to the other back-to-back inverter. This approach works well for strong-inversion operation, but it requires the bitcell to hold its data dynamically during read accesses. This approach will not work in sub-threshold because the dynamic data is susceptible to leaking away during the long access times.

The extra FETs in the 10T bitcell increase area by ~ 66% (based on layout) and also consume leakage power relative to a 6T bitcell. It is interesting to note that a 9T bitcell, identical to the bitcell in Figure 5-39 but without M_{10}, would eliminate the Read SNM problem while using less area than the 10T cell. M_{10} is valuable to the bitcell because it reduces leakage current and it allows more bitcells to share a bitline.

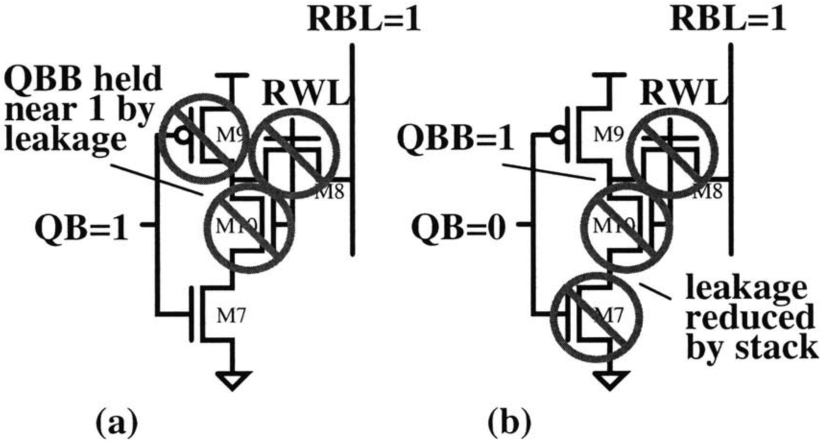


Figure 5-40: Schematic of read buffer from 10T bitcell for both data values. In both cases, leakage is reduced to the bitline and through the inverter relative to the case where M_{10} is excluded.

Figure 5-40 shows the read buffer from the 10T bitcell for Q=0 (a) and Q=1 (b). When Q=0 and QB=1 (Figure 5-40(a)), M_{10} adds an off device in series with the leakage path through M_8 and the path through M_9, decreasing the leakage through those transistors. Furthermore, since the pMOS in this 65nm technology generally has higher leakage than the nMOS, the leakage in M_9 tends to hold node QBB near V_{DD}

(see Figure 5-41), further limiting the leakage through M_8 to the bitline by making its V_{GS} negative. Even if QBB floats above 0 by only a small amount, the negative V_{GS} in M_8 reduces bitline leakage exponentially. When Q=1 and QB=0 (Figure 5-40(b)), M_{10} creates a stack of *off* nMOS transistors, reducing leakage through M_7 by the stack effect. Since node QBB is held solidly at V_{DD}, M_8 has $V_{DS} = 0$, so bitline leakage is negligible. In both cases, M_{10} reduces the leakage relative to the 9T case.

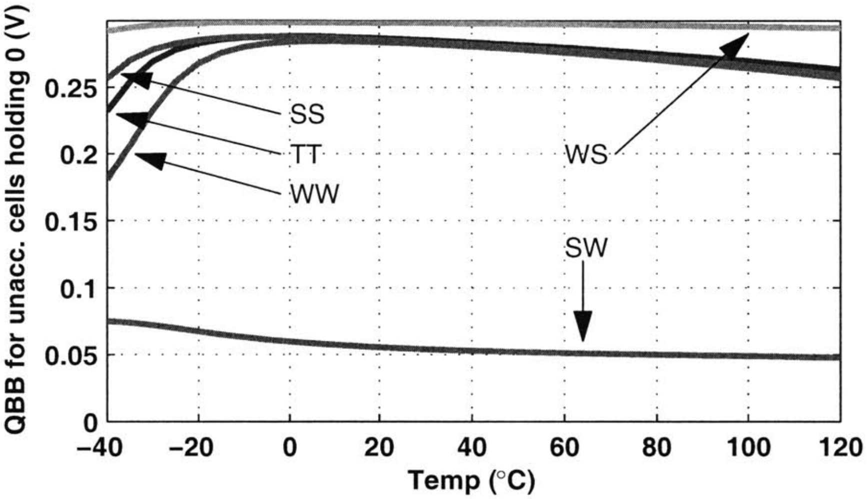


Figure 5-41: Simulation of voltage at node QBB in unaccessed 10T bitcells versus temperature and process corner. Strong pMOS leakage holds QBB near V_{DD} except at the SW corner. Even at SW, QBB is higher than it is for the 6T cell, lowering bitline leakage.

Figure 5-42 shows the relative leakage of the different bitcells under consideration. At 0.3V and nominal conditions, the 9T bitcell has 50% leakage overhead relative to the 6T bitcell. The 10T bitcell reduces this overhead to 16%. It is important to recall that while the 6T bitcell can hold data at this low voltage, it cannot function properly for either read or write accesses. Since a 6T bitcell at 600mV has the same 6σ stability as a 10T bitcell at 300mV, this overhead in leakage current is more than compensated by decreasing V_{DD} by 300mV relative to the 6T bitcell. In simulation, the 10T bitcell at 300mV consumes 2.25X less leakage power than the 6T bitcell at 0.6V (1.75X less relative to 0.5V).

The reduction in sub-threshold leakage through M_8 reduces the impact of leakage

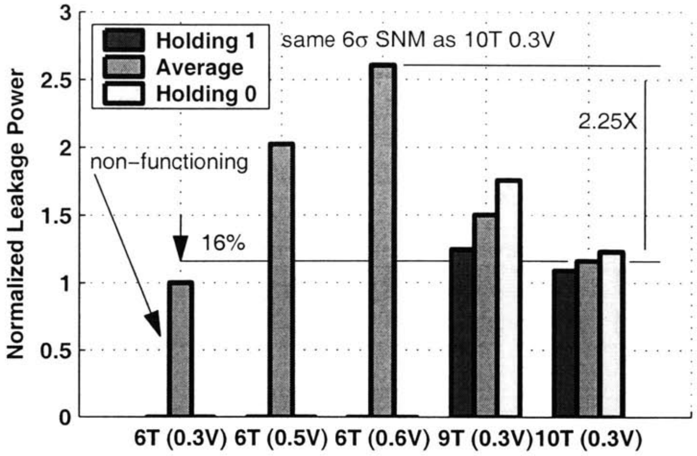


Figure 5-42: Relative leakage of bitcells. The 10T bitcell imposes 16% overhead leakage at 300mV, but the 6T bitcell cannot function at that voltage (it can only hold data). The 10T cell saves 2.25X leakage power relative to the 6T at 0.6V.

from unaccessed cells and gives the additional advantage of allowing more cells on a bitline during read. As described in Section 5.2.2, bitline leakage creates real problems for SRAMs in terms of leakage power and functionality during a read access. Leakage from the bitline into the unaccessed bitcells causes undesirable voltage changes on the bitlines. Specifically, the bitline that should remain at its precharged value of V_{DD} will droop. For differential sensing, this droop creates an effective voltage offset that the accessed cell must overcome before activating the sense amplifier, which results in longer read access times. For single ended read access like that used with the 10T cell, the steady-state voltage values for a '1' and '0' become more difficult to distinguish. Figure 5-43 shows the simulation set-up for examining the dc bitline voltages. This simulation represents the worst-case because all of the unaccessed bitcells store the complement of the data in the addressed cell. Thus, the leakage through M_{attack} counteracts the drive current of M_{cell} in the figure. The same set-up can be used to examine either 10T or 6T bitcells by using the appropriate sub-circuit in block (A). A 9T bitcell (the 10T cell without M_{10}) will look basically identical to the 6T for this experiment.

Figure 5-44 shows the impact of bitline leakage on the steady-state voltages while

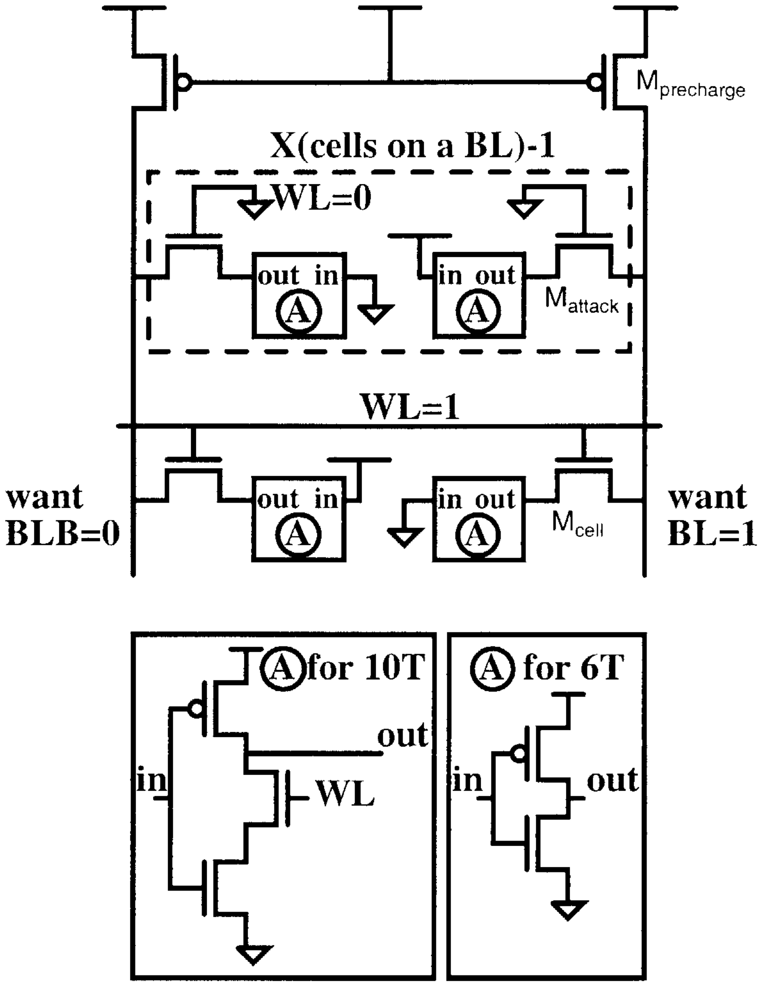


Figure 5-43: Schematic showing the simulation set-up for estimating the worst-case steady-state bitline voltages when expecting a '0' or '1'. The same setup works for both 10T and 6T depending on which cell (A) is used for the simulation.

reading a '1' (solid lines) or '0' (dotted lines). For the same number of cells on a BL, the 10T bitcell (circles) shows larger bitline separation than the 6T (or 9T) bitcells (squares). This figure suggests that 'sensing' with an inverter (whose switching threshold, V_M, is shown) should work well from 0°C to 100°C even with 256 cells on a bitline for the 10T cell. In contrast, as previously discussed in Section 5.2.2, the 6T cell (or 9T bitcell) would allow at most 16 bitcells on a bitline. The bitline that should be '1' stays very close to V_{DD} at high temperatures and then begins to

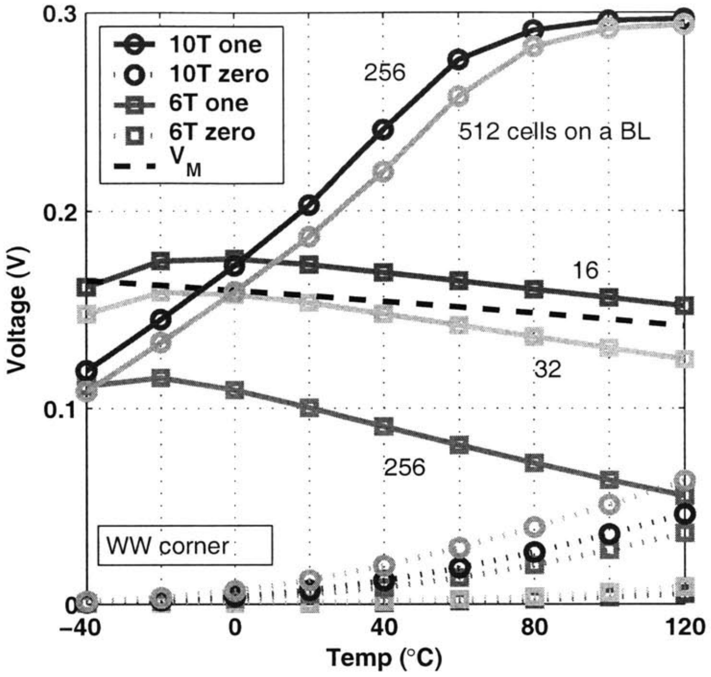


Figure 5-44: Simulation of scenario in Figure 5-43 showing steady-state bitline voltages. The 10T bitcell exhibits much better steady-state bitline separation than the 6T cell. The WW corner is shown at 300mV.

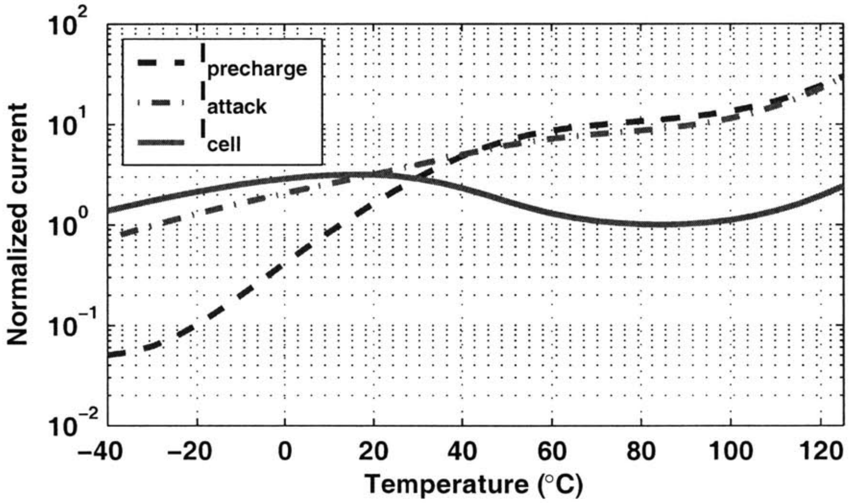


Figure 5-45: Plot showing relative currents at steady-state in the worst-case for bitline leakage for the 10T cell (transistors in Figure 5-43) at the WW corner.

droop at lower temperatures. Figure 5-45 gives a hint as to why this occurs. M_{10} inside the unaccessed 10T bitcells is so successful at reducing sub-threshold current through the access transistors that the sub-threshold current actually drops below the gate leakage (which is fairly constant with temperature). At higher temperatures, the leakage through the pMOS precharge device, $M_{precharge}$ in Figure 5-43, exceeds the gate leakage into the unaccessed cells (M_{attack} in Figure 5-43) and holds the bitline close to V_{DD}. As temperature decreases, the sub-threshold leakage through $M_{precharge}$ drops faster than the sum of gate leakage into M_{attack} until the accessed cell begins to take over. At this point, the bitline voltage has to droop in order for M_{cell} and $M_{precharge}$ to supply the current that flows into M_{attack}. If gate leakage was lower (perhaps in the case of high-K dielectrics), then sub-threshold leakage into the unaccessed cells is reduced sufficiently such that the bitline will stay very close to V_{DD}.

Figure 5-46 shows the comparison of 10T and 6T steady-state bitline voltages at the remaining process corners (SW in (a), TT corner in (b), WS corner in (c), and SS corner in (d)). The only corner that differs noticeably from the WW corner in Figure 5-44 is the SW corner in Figure 5-46(a). The bitline leakage into the unaccessed bitcells is not reduced by nearly as much at this process corner. Since the QBB node in the 10T bitcell is held high by leakage current, its voltage value depends on the relative leakage of the transistors inside the bitcell. As clear from Figure 5-41, the stronger nMOS leakage at this corner keeps QBB closer to 0V, allowing more sub-threshold current to flow from the bitline. This increased bitline leakage degrades the voltage at the bitline holding a '1', although the 10T cell still has more than double the dc bitline separation of the 6T cell at higher temperatures. Figure 5-46 makes it clear that the 10T bitcell allows a much higher number of bitcells on the bitline than the 6T bitcell across all of the process corners.

5.4.2 Enabling Sub-threshold Write

Write functionality offers the second primary obstacle to sub-threshold SRAM. In this 65nm technology, a 6T bitcell cannot write in the traditional fashion below around

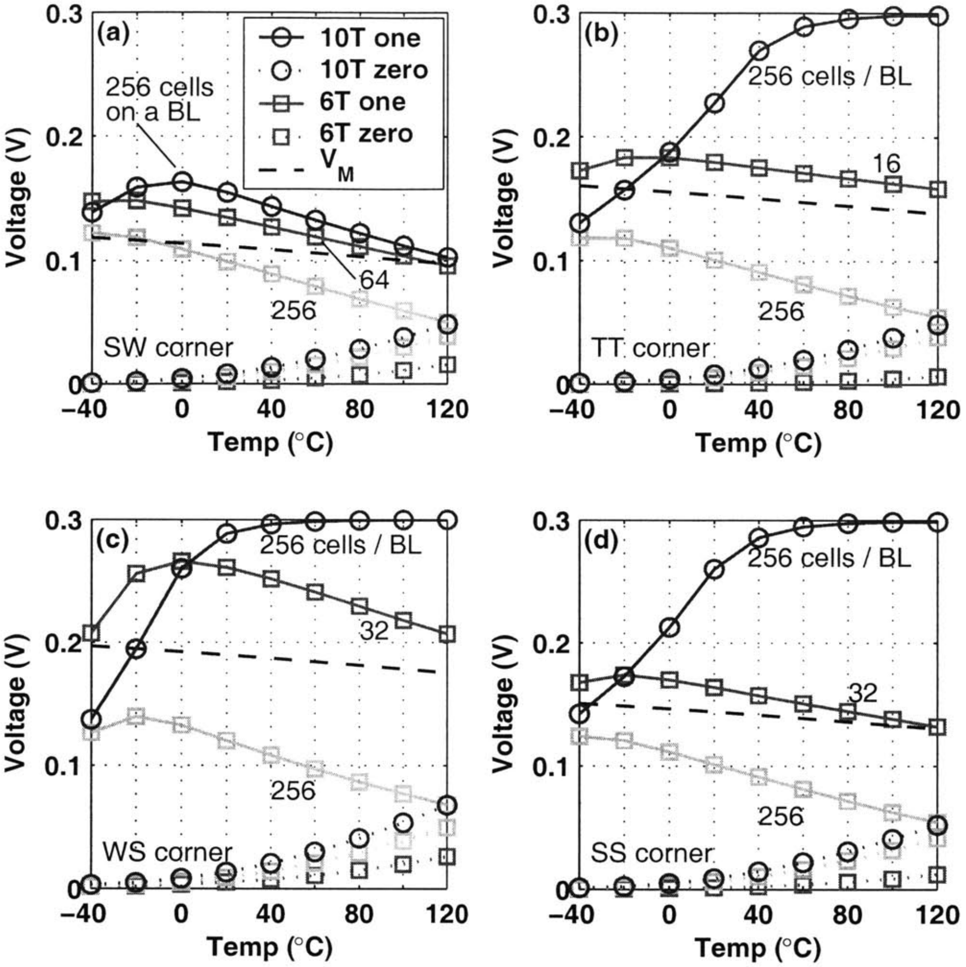


Figure 5-46: Simulation of scenario in Figure 5-43 showing steady-state bitline voltages for the SW corner (a), the TT corner (b), the WS corner, (c), and the SS corner (d).

0.6V, as we described in Section 5.2.1. The primary reason for write failure was the inability of the write driver and nMOS access transistor to win the ratioed fight against the pMOS inside the bitcell and to write a '0'. Section 5.1.3 described previous works that use a virtual supply rail that floats during a write access. Although those works applied this approach primarily for increasing speed, the method itself addresses the problem that sub-threshold bitcells face.

Since the pMOS devices in the 10T bitcell are the problem, our approach uses a

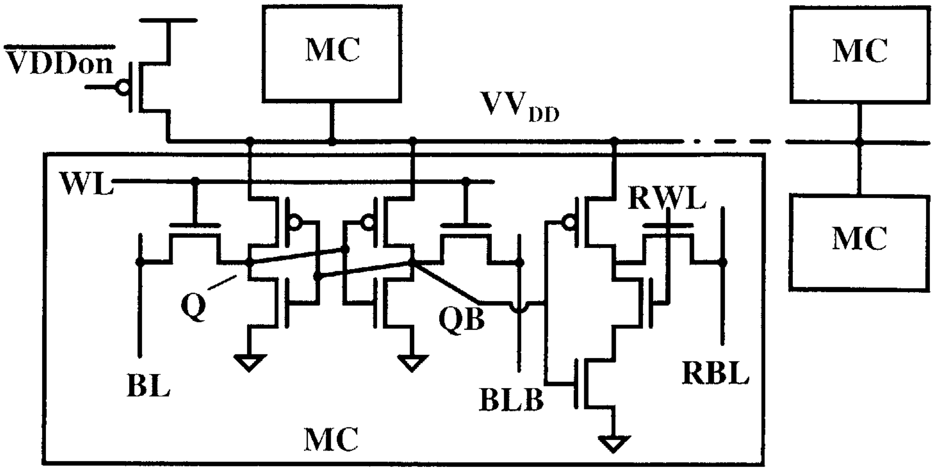


Figure 5-47: Schematic of write architecture for a single row using a floating power supply (VV_{DD}). The row is 'folded' in layout so that its cells share n-wells, and the entire row is written at once.

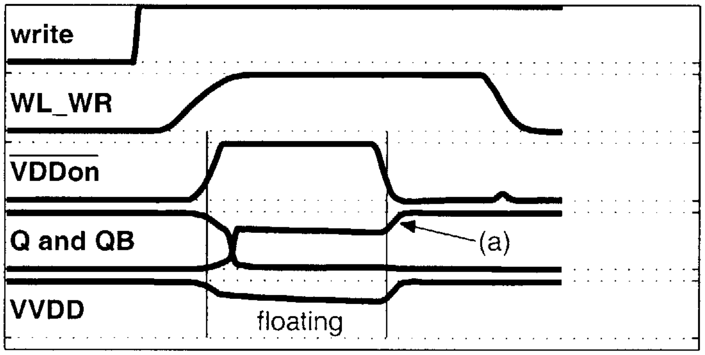


Figure 5-48: Timing diagram for write operation. When $\overline{V_{DDon}}$ goes low while WL_{WR} remains asserted, the cell's feedback restores full voltage levels for the new values of Q and QB (point (a)).

virtual power supply rail rather than a virtual ground rail. Figure 5-47 shows the simple schematic for a single row using this approach. A single power-supply-gating header switch connects node VV_{DD} to the true power rail. When the bitcell holds its data or during read accesses, $\overline{V_{DDon}} = 0$ so that $VV_{DD} = V_{DD}$. During a write access, the virtual rail floats.

Figure 5-48 shows the timing associated with a write access using this scheme. First, the 'write' signal goes high to indicate that a write access will occur, and the bitlines (BL and BLB in Figure 5-47) are driven with the new data. Next, the decoders drive a global wordline (not shown) which eventually causes the local write wordline (WL_{WR}) to go high. Triggered by the local wordline, the $\overline{V_{DDon}}$ signal goes high, allowing node VV_{DD} to float. As the write access transistors discharge the virtual rail, its voltage droops, and Q and QB change to their new values. While VV_{DD} continues to float, denoted by the 'floating' label on the timing diagram, the logical '1' inside the cell tracks its drooping voltage value. When $\overline{V_{DDon}}$ goes low again while the *local wordline remains high*, it reconnects the virtual rail to the full supply. The feedback inside the bitcell then holds the Q and QB nodes at their correct logical values and amplifies the '1' to full V_{DD}. This occurs at point (a) in Figure 5-48.

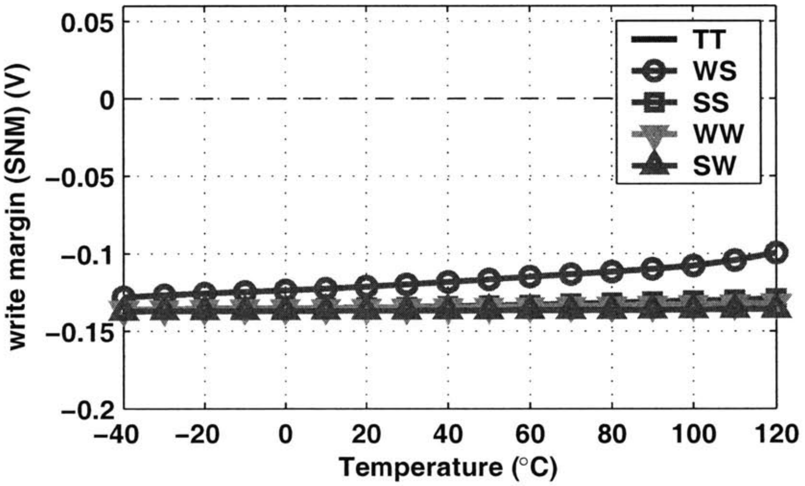


Figure 5-49: Write margin (write SNM) versus temperature at 0.3V for 10T bitcell with floating VV_{DD} supply. Negative margin for all corners, signifying successful write operation.

Some previous works implement a floating rail in the column-wise direction. The risk of the column-wise approach is that any droop that occurs on VV_{DD} during a write operation will impact other unaccessed bitcells that are holding their data. For sub-threshold operation, the lack of voltage headroom increases the risk of losing data in those cells by decreasing their Hold SNM. For this reason, we implement the virtual rail along a row of the memory, as Figure 5-47 shows. For the implementation on the test chip, a conceptual row is folded as shown in the figure so that its bitcells can share n-wells, and the entire row is written at once.

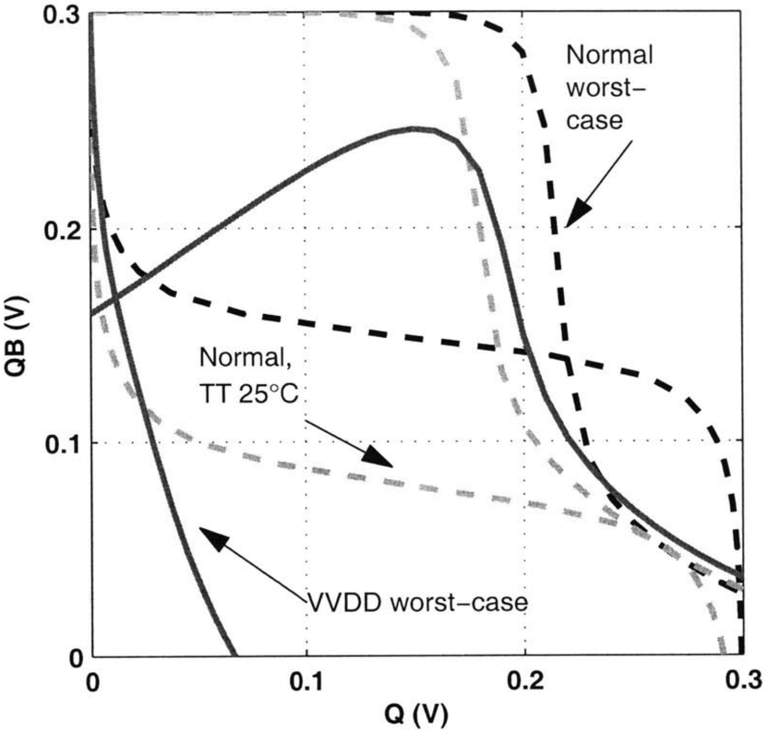


Figure 5-50: VTCs showing write SNM with and without using a virtual power rail (VV_{DD}) at $V_{DD} = 300$mV. Normal write fails even under normal conditions of process corner and temperature, but the virtual rail approach works even in the worst-case as shown by significant negative noise margin.

The plot in Figure 5-49 shows the write margin for the virtual V_{DD} approach across temperature and process corner at $V_{DD} = 300$mV. The write margin remains negative across all of these ranges, indicating a successful write. The worst-case write margin occurs at the WS corner and high temperature. Figure 5-50 overlays the VTCs at this worst-case point on the VTCs for a 'normal' 6T bitcell during write at the worst-case and typical corners. In both cases, the write fails for the 6T bitcell as evidenced by continued bistability in the cell. The VTCs for the 10T bitcell using the virtual rail approach show clear mono-stability. For the 10T cell in this example, a '0' is written to node Q, and '1' is written to node QB. The logical '1' is degraded below V_{DD} because of the droop on the VV_{DD} node. As illustrated in Figure 5-48, this voltage will recover to full V_{DD} when $\overline{V_{DDon}}$ goes low again and the wordline stays asserted. The plot shows that, even for the worst-case, this method provides ample negative noise margin for ensuring a write.

Since the 10T bitcell shows the ability to solve both the read and write problems for sub-threshold operation, we chose it as the bitcell for a test chip in 65nm bulk CMOS.

5.5 65nm Sub-threshold SRAM Test Chip

This section describes the test chip that uses the 10T bitcell to enable sub-threshold operation.

5.5.1 Test Chip Architecture

A 256kb 65nm bulk CMOS test chip uses the 10T bitcell and the architecture shown in Figure 5-51. The memory is divided into eight 32-kb blocks. Each block contains an array of 256 rows and 128 columns of 10T bitcells. A single 128-bit Data Input/Output (DIO) bus serves all eight blocks. In this initial instantiation of the sub-threshold memory, only one read or write can occur per cycle. As mentioned previously, however, the 10T bitcell can accommodate both a read and write access to the same block in a single cycle. Such a dual-port instantiation of the memory would require a second DIO bus and additional peripheral logic. The decoder in this memory uses the top three address bits to determine the block and generates a block select signal (BKsel) to enable certain local features within the selected block. The remaining eight address bits select the correct row inside the block. The decoder decodes these eight bits and asserts a global wordline. The global wordline then asserts a local wordline inside the selected block. The local wordline then combines with the local write signal to assert either WL_{RD} or WL_{WR}. For a write access, local logic turns off $M_P\langle\mathrm{r}\rangle$ to the accessed row as described in Section 5.4.2. The write drivers consist simply of inverters with transmission gates. Although unnecessary for functionality in this design, the transmission gates turn off when the memory is not

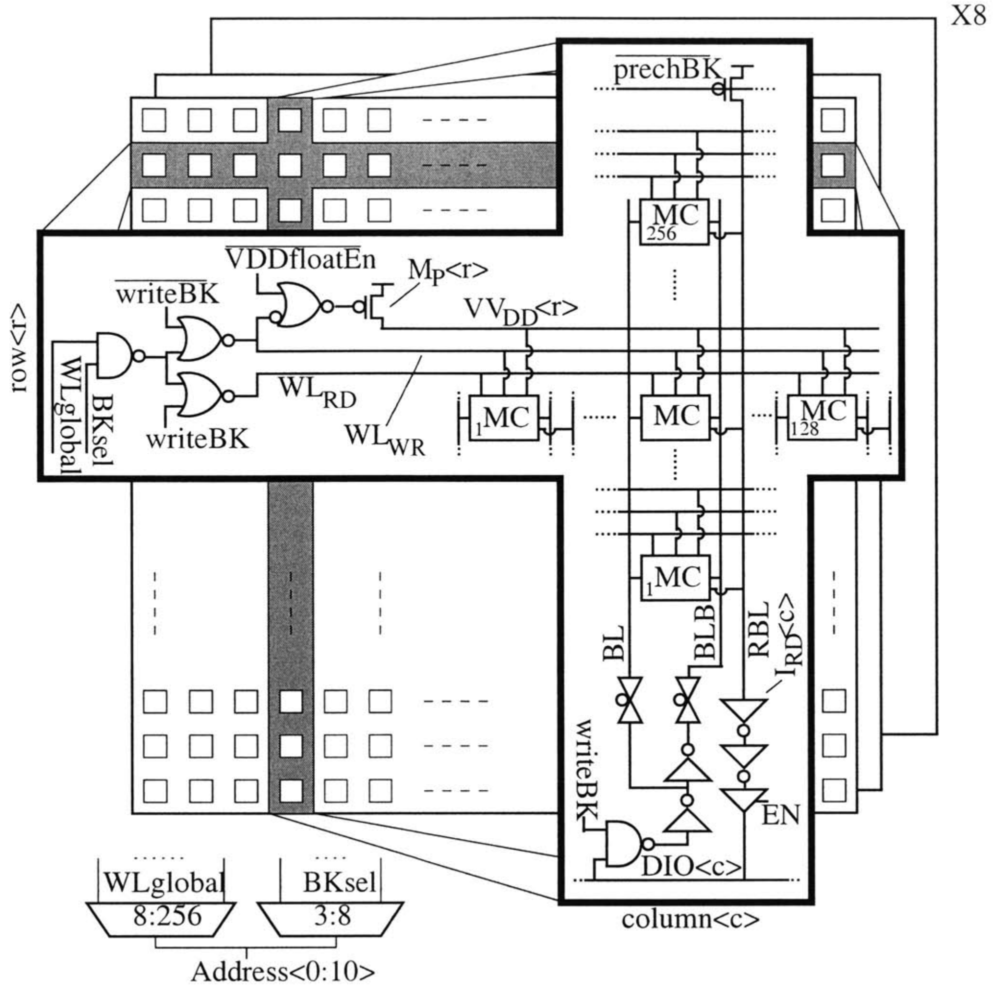


Figure 5-51: Architecture diagram of the 256kb memory on the test chip using 10T sub-threshold bitcells.

writing to minimize leakage on the write bitlines (BL and BLB). The power supply to the WL drivers is routed separately to allow a boosted WL voltage. This technique improves the access speed and increases the robustness to local variations. The read bitline (RBL) is precharged prior to read access, and its steady-state value is 'sensed' using a simple inverter, $I_{RD}\langle c \rangle$, as discussed in Section 5.4.1. Tristate buffers prevent the output of the blocks from driving the DIO bus at the incorrect times. Column and row redundancy is a ubiquitous technique in commercial memories used to improve yield. For our analysis of the SRAM, we assume the availability of one redundant row and column per block.

The primary goals for this test chip were to test the functionality of the 10T bitcell in sub-threshold and to explore the limitations of the design. For this reason, the peripheral circuits were designed to be as simple as possible. All of the peripherals use static CMOS logic for simplicity and for functional robustness in sub-threshold. The large block size was intentionally aggressive in order to expose limitations in the bitcell and architecture. Integrating 256 bitcells on the bitline (as opposed to 16 for 6T) pushes the envelope for functionality.

The layout of the memory had to meet logic design rules. For this reason, even the reference 6T bitcell layout is much larger than commercial 6T layouts for which the design rules are relaxed. The 10T bitcell layout added almost exactly the expected 66% area overhead relative to our reference 6T design. The area penalty for the entire array may not be so large, however. Since the 10T bitcell allows 256 bitcells on a bitline, fewer copies of the column-wise peripherals are necessary than for the 6T cell array, which can only accommodate 16 cells per bitline pair. As an example, the 8T bitcell in [124] has 40% larger area than its 6T counterpart. However, by equalizing bitline leakage (see Section 5.1.4), the bitcell allows 256 cells per bitline rather than 16. The total cache using the 8T bitcell in a 100nm technology ends up being smaller than its 6T counterpart by 6% [124]. This example suggests that the total array area penalty for the 10T cell is much less than 66%, since it gives a similar advantage in bitline integration.

Table 5.3: 6T and 10T architecture comparison.

	6T		**10T**	
line	**number**	**transistors**	**number**	**transistors**
write WL	1	256	1	256
write BL	2	256	2	256
read WL	1	256	1	256
read BL	2	256	1	256

As we described in Section 5.4.2, we chose for this implementation to switch the n-wells along a row along with VV_{DD}. This approach made it easier to follow the design rules related to distance between well taps and avoided the need to route an additional V_{DD} rail. To make this approach work, each row is folded such that a pair of 64-bit physical rows sharing n-wells and a VV_{DD} rail makes up one conceptual 128-bit row (c.f. Figure 5-47). This folding increases the length of bitlines by roughly 2X and decreases the length of wordlines by roughly $\frac{1}{2}$X. Notice that this is not fundamentally necessary for the write approach to work. The n-wells of two separate rows can be shared and the VV_{DD} for each row routed separately.

The impact of the 10T approach on the number of wordlines and bitlines used during memory accesses is beneficial. Table 5.3 shows the comparison for a 6T memory that also has 256 rows and 128 columns (this could not actually function in sub-threshold or even above threshold because of bitline leakage). The 10T approach uses one less bitline, and the transistor load on any given wordline or bitline is the same. As we mentioned previously, separate wordlines and bitlines for write and read accesses allow simultaneous write and read accesses to the memory.

Figure 5-52 shows a layout shot and die photograph of the test chip. The die size is 1.89mm by 1.12mm, and the chip is pin-limited. The 256kb array and a 32kb block are highlighted for reference. Metal fill in all of the metal layers obscures the features in the interior of the die photograph.

5.5.2 Measurements

Measurements of the SRAM test chip confirm that it is functional over a range of voltages from 1.2V down into the sub-threshold region. With the assumption of one redundant row and column per block, the memory operates correctly to below 400mV. Read operation works without error to 320mV and write operation works without error to 380mV at 27°C. We continued to push the supply voltage to even lower values to examine the limits of the implementation. At the low supply voltage of 300mV, the memory continues to function, but it does exhibit bit errors in $\sim 1\%$ of its bits that result from sensitivities in the architecture to local device variation. Figure 5-53 shows an oscilloscope plot of two data bits output from the memory during read operations at 300mV. Later in this section, we discuss the specific causes of the bit errors, and Section 5.5.3 proposes improvements to the architecture that will eliminate them. Although the bit errors indicate that the issue of memory yield requires further attention, the test chip successfully demonstrates a functional sub-threshold memory that overcomes the problems it was designed to face.

First, the bitcell removes the Read SNM problem. Measurements have confirmed that the memory experiences zero destructive read errors at 300mV. Simulations show that a 6T memory would experience a high rate of destructive read errors at 300mV due to degraded Read SNM. Secondly, whereas a 6T memory would fail to write below

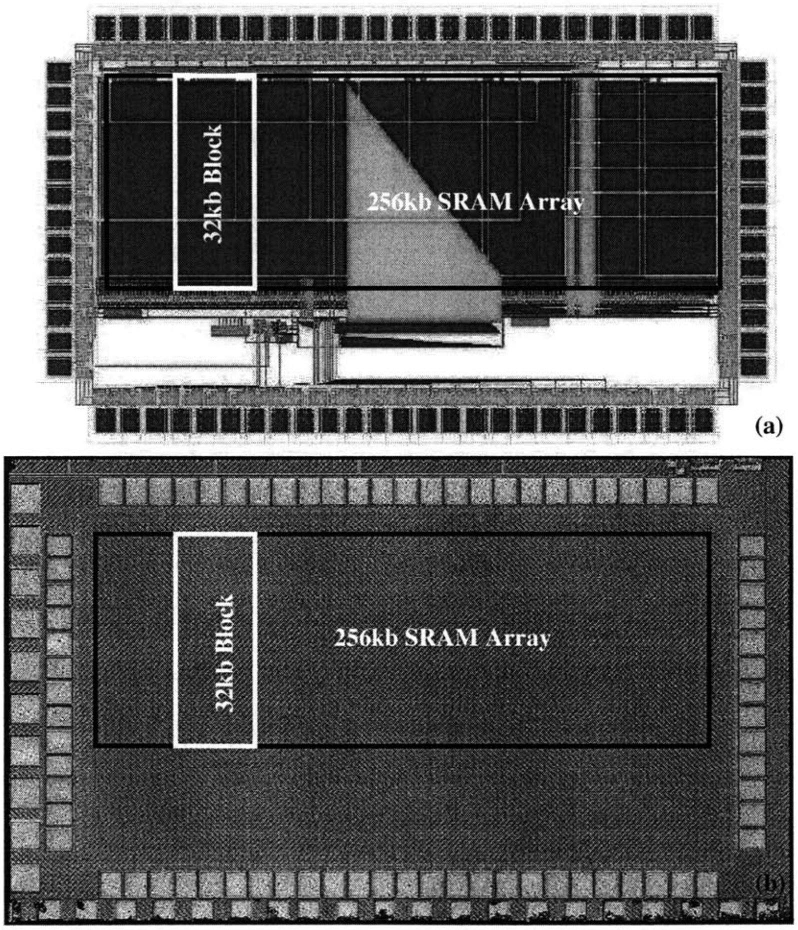


Figure 5-52: Annotated layout (a) and die photograph (b) of the 256kb sub-threshold SRAM in 65nm. Die size is 1.89mm by 1.12mm.

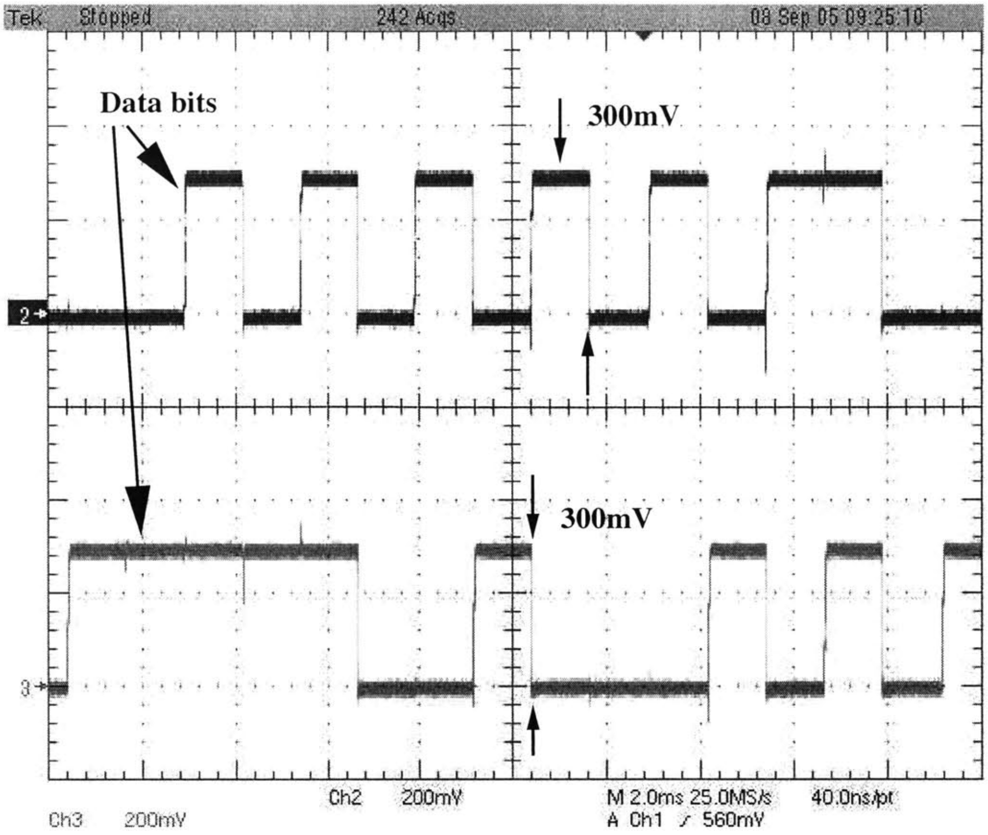


Figure 5-53: Oscilloscope waveform showing correct functionality at $V_{DD} = 300\text{mV}$. At this low voltage, a small fraction of bits have errors.

about 600mV, this memory writes correctly at 350mV at 85°C. Thirdly, a 6T memory would experience problems reading with only 16 bitcells on a bitline. Measurements show that the 10T memory reads correctly even with 256 bitcells on the bitline down to 320mV. Finally, the memory shows good Hold SNM performance. The first bits observed to fail to hold their data occur at $V_{DD} < 250\text{mV}$. An analysis of the bit errors at lower voltages indicates that changes to the periphery circuits of the array should eliminate them.

Figure 5-54 shows the measured leakage power of the test chip at two different temperatures. As expected, voltage scaling provides significant reduction in leakage power. Figure 5-55 shows the relative savings in leakage power from V_{DD} scaling. The plot is normalized to operation at 0.6V for comparison to a theoretical minimum-

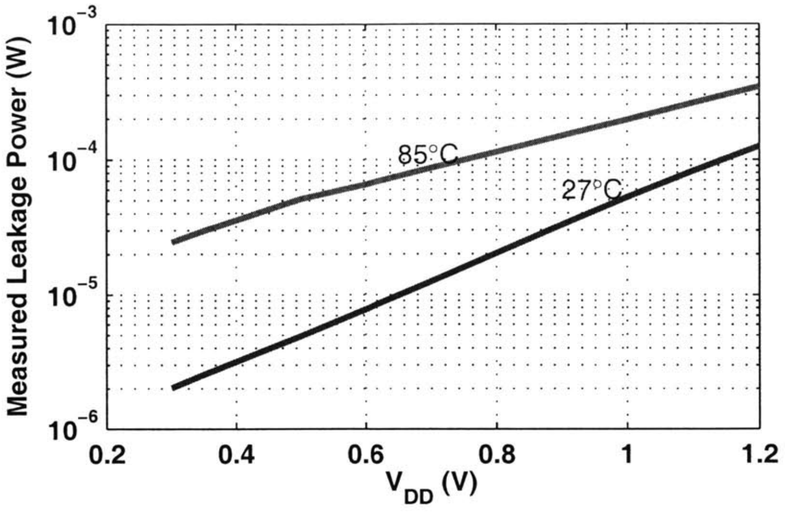


Figure 5-54: Measured leakage power from the memory test chip.

voltage 6T memory. At 27°C, the 10T memory saves 2.5X and 3.8X in leakage power by scaling from 0.6V to 0.4V and 0.3V, respectively. Leakage power decreases by over 60X when V_{DD} scales from 1.2V to 0.3V. Voltage supply scaling also gives the expected savings in active energy. Figure 5-56 shows the energy per read access versus supply voltage based on the measured switched capacitance of the memory.

Read Bit Errors

As we described above, both read and write operations at low voltage expose bit errors in a small fraction of the bits. The errors are deterministic in the sense that they occur for the same set of bits in a repeatable fashion. The read bit errors are all of the same nature. For certain bits at low voltage, a read access shows that the bitcell holds a '0' when in fact it holds a '1'. This error is non-destructive, which we show by raising the supply voltage and re-reading the cell. Invariably, this check provides the correct value of '1'. Additionally, these bit errors tend to be gathered along a small number of specific columns. We have shown that the error is a steady-state problem by extending the read cycle time to well beyond any transients in the read operation. Thus, the mechanism for error must explain why the memory cannot

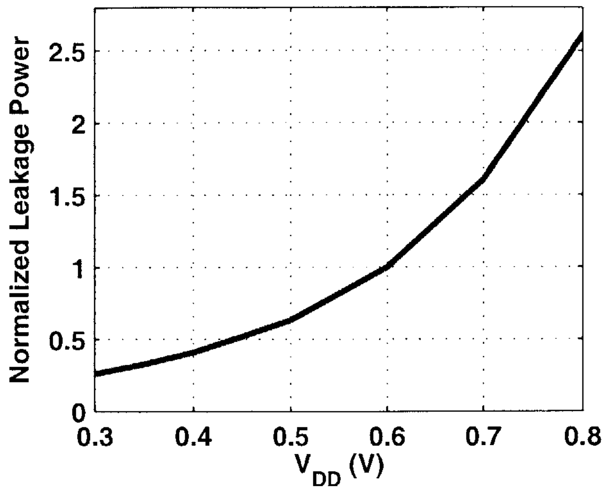


Figure 5-55: Relative leakage power savings at 27°C achieved by V_{DD} scaling.

detect the stored '1' inside of the cell. No failures to read a '0' have been observed.

The fact that this error exists in a small fraction of cases indicates that it results from local device variation, and we can isolate the problematic transistors. The fact that the bits exhibiting problems cluster along specific columns indicates that variation in the sensing inverter, I_{RD} in Figure 5-51, has shifted its switching threshold,

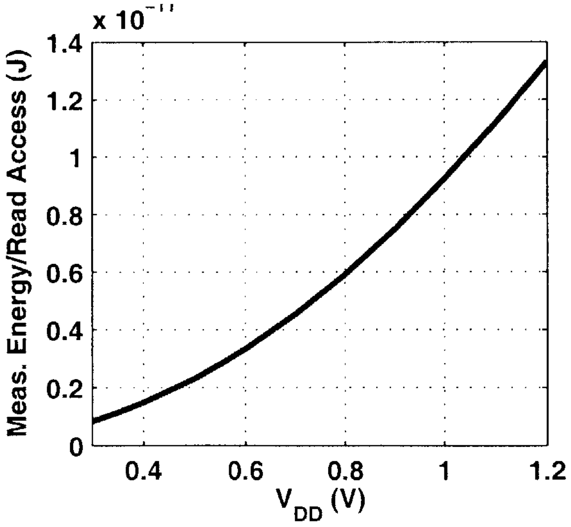


Figure 5-56: Measured active energy per read access.

V_M, towards V_{DD}. Now, specific bitcells along this column that have read access transistors weakened by local variation cannot hold the read bitline above V_M of inverter I_{RD}. Several experiments confirm that this is the mechanism for read failures. First, we can independently lower the supply voltage of the sense inverters, $I_{RD}\langle c \rangle$. This lowers the V_M for the inverters, and the measured bit error rate for read decreases. Secondly, we can increase temperature. The plots in Figure 5-46 show that the bitline separation should improve at higher temperature. Measurements show the same trend, although the improvement is not as dramatic as Figure 5-46 suggests it will be. Finally, we can increase the voltage of the wordline drivers. The larger V_{GS} for the access transistors makes them stronger, pulling the transistors that weakened by local variation back toward the mean. The result is that bit errors decrease rapidly with increased wordline voltage.

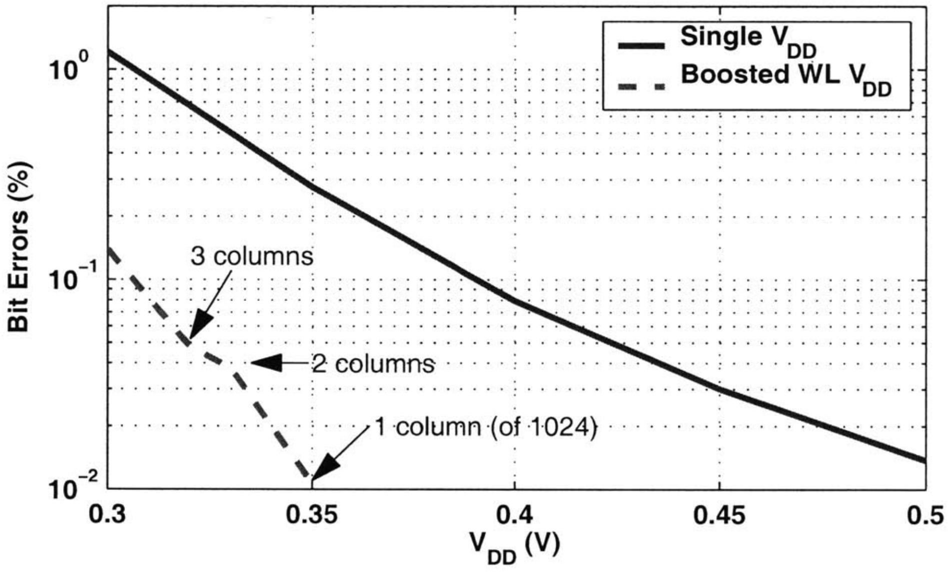


Figure 5-57: Measured percentage of bit errors for read versus V_{DD}. Boosting the WL voltage dramatically reduces these errors.

Figure 5-57 shows the measured percentage of bit errors during read access versus supply voltage. The error rate without the wordline boosted by 100mV also is shown. Wordline boosting is common in DRAM and has been used in SRAMs to increase access speed (e.g. [94]). It also serves well as a mechanism for increasing reliability in sub-threshold. In sub-threshold, the extra gate voltage ($\sim$ 100mV) on the read access transistors provides roughly one order of magnitude (due to the sub-threshold slope) of extra current drive. As with above-threshold memories, this extra drive current provides faster operation. It also makes the design more robust to mismatch.

By aggressively choosing a block having 256 rows on a single bitline, we pushed the limits of read operation and exposed bit errors that result from local variation. Boosting the wordline voltage offers one simple change that dramatically reduces the error rate. In Section 5.5.3, we discuss changes to the design of the memory to eliminate this type of read error. Even without changing the architecture, we can combine boosted wordlines with column redundancy, which is a ubiquitous approach to improving SRAM yield (e.g. [146]). Assuming only one redundant column per block, this memory can read without error at 320mV.

Write Bit Errors

Write errors also occur at very low supply voltages. These errors appear as the failure to write specific bitcells in a deterministic way. Bit errors are observed for writing both data values. The write errors tend to aggregate in bits along specific rows. As with the read errors, the small number of errors points to local device variation, and the predominance of row-wise errors suggests that the failure mechanism involves the row peripherals.

Referring back to Figure 5-51, write operation fails along specific rows whose pull-up device, M_P, is strengthened by local variation. Thus, when M_P turns off during a write access, it has stronger leakage that makes it harder for the bitcells to increase the voltage droop on VV_{DD}. Some of the bitcells can still switch under these conditions, but the voltage on VV_{DD} reaches a steady-state value that is high enough to prevent some bitcells from overpowering the pMOS to write a '0' into the memory. In these bitcells, local mismatch has made the internal pMOS relatively stronger than the access transistor to the point that the write driver cannot flip the cell at the steady-state VV_{DD} voltage.

Measurements confirm that this is the case. First, for this 65nm process, pMOS leakage reduces relative to nMOS leakage as the temperature increases. Measurements confirm that the lowest functional supply voltage decreases at higher temperature. Since the leakage through M_P gets relatively weaker compared with the nMOS access transistors, this confirms the mechanism for failure. More importantly, the write errors decrease when the supply voltage to the wordline increases. As with read accesses, the higher wordline voltage increases V_{GS} for the write access transistors and makes them more capable of producing voltage droop on VV_{DD}.

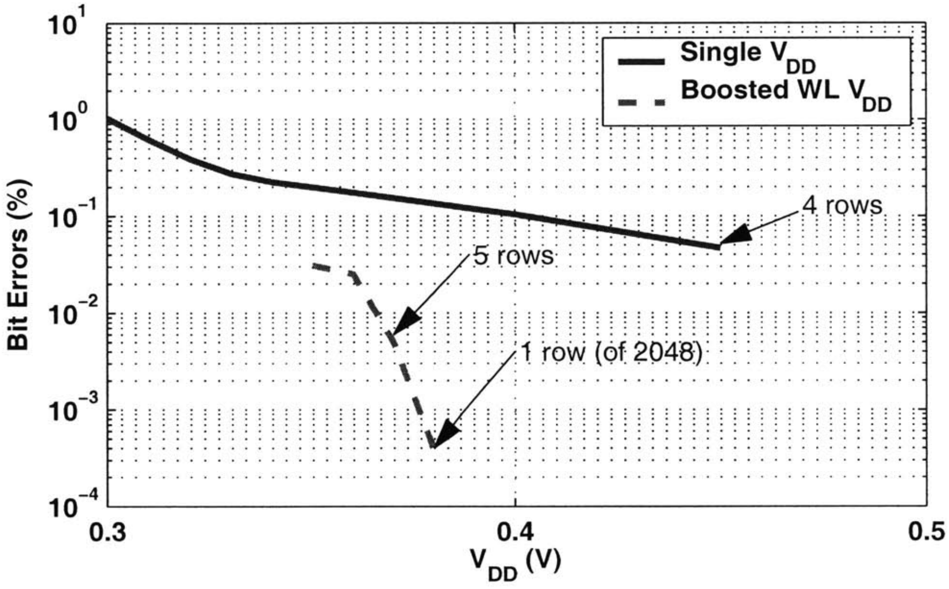


Figure 5-58: Measured percentage of bit errors for write versus V_{DD}. Again, boosting the WL voltage dramatically reduces these errors.

Figure 5-58 shows the percentage of bit errors measured during write both with and without 100mV of wordline boosting. Wordline boosting improves the bit error rate significantly. Also, the errors occur along a small percentage of the overall rows as shown. Assuming one redundant row per block, the memory can write without error at 380mV at 27^{o}C and 350mV at 85^{o}C.

5.5.3 Proposed Improvements

The previous analysis of bit errors showed that the commonly used techniques of wordline boosting and row/column redundancy go a long way towards removing bit errors even at very low voltages. With only one redundant row and column per block ($< 1\%$ redundancy) and 100mV wordline boosting, the memory functions to below 400mV. More redundancy can push this even lower. However, it is preferable to correct the bit errors at a more fundamental level. The 10T bitcell itself works quite well, and the observed problems with the test chip reside in the peripheral circuits. Changing these circuits to make them more robust to local variation by design will improve the yield for the sub-threshold memory.

Improving Read

The read bit errors come from the inability to sense the correct value when reading a '1'. The most obvious problem is using an inverter as an overly simple sense amplifier. Some improvements are possible even if the inverter is retained as the sensing circuit. The switching threshold of each column could be tuned on a column-by-column basis, for example by adjusting the effective size of the either the nMOS or pMOS by using different numbers of parallel devices. Other more complicated methods of biasing could also be used. This type of local tuning is not at all uncommon in modern processors. For example, local clock buffers are individually tuned to balance skew.

The better solution to improve the read reliability and robustness to local device variation is to replace the inverter with a new sensing scheme. As the bitline separation plots have shown, V_M of the sensing inverter lies too close to the logical '1' value at some corners and temperatures. The literature is replete with a variety of sensing architectures that can do a better job of distinguishing between '0' and '1'. Approaches to DRAM sensing that take an inherently single-ended bitcell and convert to pseudo-differential sensing are quite applicable. Improving the sensing scheme is the most promising approach to removing bit errors during read.

Improving Write

The main problem with writing is that the voltage droop on VV_{DD} does not develop sufficiently for the weakest cells on the row to switch the cell data. The leakage through the VV_{DD} pull-up switch is too strong on some rows because of device variation. Several well-known methods are available that, when applied to M_P, could remedy this issue. For example, implementing M_P as a stack of multiple series transistors will lower its leakage current. Likewise, applying RBB directly to M_P has the same effect. Since M_P is a pMOS, a triple well process is unnecessary. As we described before it is not fundamentally necessary to tie the n-wells along a row to VV_{DD}. Instead, at the cost of routing a separate V_{DD} to tap to the n-wells, the n-wells of all rows can stay at V_{DD} even during write access. Then, the row whose virtual rail floats will experience RBB in its pMOS transistors. This allows the nMOS access transistors to overpower the bitcell feedback in the presence of less voltage droop on VV_{DD}.

A better solution to the write issue that maintains the same basic architecture and approach is to induce a specific voltage drop on VV_{DD} intentionally. In the extreme, replacing M_P with an inverter will drive VV_{DD} all the way to 0V. Then, as long as the write wordline remains asserted, the bitcells will develop the correct internal data when VV_{DD} goes back high regardless of local variations. A disadvantage of this extreme case is the energy penalty associated with discharging and re-powering the VV_{DD} rail and all of the bitcells in the row. An alternative is to mimic some of the previous works from Section 5.1.3 and use a circuit (e.g. diode connected FET) to force VV_{DD} to some intermediate value that is low enough to ensure write but that uses less energy.

5.6 Summary and Conclusions

In summary, sub-threshold SRAM provides the dual advantages of minimizing total memory energy consumption and of providing compatability with minimum-energy sub-threshold logic. Traditional 6T SRAM cannot function in sub-threshold because it fails to write below $\sim$ 600mV and because the Read SNM degrades badly at low supply voltage. Furthermore, bitline leakage in 6T SRAMs limits the number of bitcells on a bitline to around 16. A 10T bitcell solves these problems and provides functionality into the sub-threshold region. The bitcell solves the write problem by using a floating supply voltage that allows the write drivers to overcome the cell feedback. A read buffer prevents read accesses from affecting the stored data and thus removes the Read SNM problem. The read buffer uses a low-leakage design that allows many more cells to use the same bitline relative to the 6T bitcell.

A 256kb 65nm bulk CMOS test chip uses the 10T bitcell and demonstrates sub-threshold operation. The memory functions correctly down to 300mV, although bit errors appear in around 1% of the bits at this low voltage. Although aggressive design exposes the limitations of the architecture in terms of its robustness to local device variation, the bit errors result primarily from problems in the peripheral circuits. Measurements show that the bitcell fundamentally solves the Read SNM problem, overcomes the write problem, and relaxes the bitline integration limitation. With one redundant row and column per block and a boosted wordline, the memory functions without error to below 400mV. At 400mV, it consumes 3.28mW and works up to 475kHz. No bit errors for holding data occur in the SRAM until V_{DD} scales below 250mV.

Chapter 6

Conclusions

Theoretical treatment of sub-threshold operation for digital circuits prior to the work presented in this thesis makes the case that sub-threshold circuits are valuable for low energy scenarios. The FFT processor presented in [4] showed a full sub-threshold digital system whose minimum energy point occurs in the sub-threshold region. This result paves the way for more research into sub-threshold circuits with the eventual goal of empowering designers to use sub-threshold circuits for low-energy designs. This thesis contributes to achieving this goal in four key areas.

6.1 Summary of Contributions

Modeling and Characterization

- Analytical solution for the optimum V_{DD} and V_T to minimize energy for a given frequency for arbitrary circuits in the sub-threshold region.

- Demonstrated and characterized minimum energy point dependencies on the technology, the characteristics of the design, and on operating conditions such as temperature, duty cycle, and workload.

- Simulations and measurements from a 0.18μm CMOS programmable FIR test chip matched the model within a few percent.

Sizing Analysis

- Minimum sized devices are theoretically optimal for minimizing power.
- Linear impact of sizing on current makes it a less effective knob for compensating for V_T differences that have an exponential impact on current in sub-threshold.
- Measurements from a 0.18μm CMOS programmable FIR test chip confirm that existing static CMOS standard cell libraries function well in sub-threshold.
- A standard cell library primarily using minimum-sized devices would theoretically minimize energy per operation.

Ultra-Dynamic Voltage Scaling

- Low overhead method of Local Voltage Dithering (LVD) for frequency and voltage scaling across a large range while minimizing energy.
- Ultra-Dynamic Voltage Scaling (UDVS) combines LVD with sub-threshold operation to offer frequency and voltage scaling across the full operational range of a circuit with near-optimum energy consumption.
- A 90nm CMOS test chip provides measurements for the analysis verifying the energy savings achievable by LVD and UDVS.
- Data from the test chip shows that both the transition delay and the energy overhead associated with LVD are relatively small.
- Using UDVS to scale to the optimum supply voltage reduces total energy consumption by 9X.

Sub-threshold SRAM

- Sub-threshold SRAM provides the dual advantages of minimizing total memory energy consumption and of providing compatability with minimum-energy sub-threshold logic.

- 6T SRAM bitcell and architecture cannot function below 0.6-0.7V due to fundamental problems with write failure and Read SNM.

- Bitline leakage in 6T SRAMs limits the number of bitcells on a bitline to around 16.

- A 10T bitcell designed to function in sub-threshold solves the write problem and removes the Read SNM problem.

- Reduced bitline leakage allows many more bitcells on a bitline than 6T.

- A 256kb 65nm bulk CMOS test chip uses the 10T bitcell and demonstrates sub-threshold operation.

- Assuming one redundant row and column per block, the memory functions without error to below 400mV. At 400mV, it consumes 3.28μW and works up to 475kHz.

- No bit errors for holding data occur in the SRAM until V_{DD} scales below 250mV.

6.2 Conclusions and Open Problems

The work that this thesis describes advances the field of sub-threshold design. Results in the areas of modeling, sizing, full-range voltage scaling, and memory provide methods for designers to integrate into complete sub-threshold systems. This section offers the key lessons and conclusions from each of these areas and discusses opportunities for future work.

The relatively simple model we present for analyzing the minimum energy point demonstrates the predictable nature of the supply voltage that minimizes energy. Despite the changes of delay and leakage current of potentially orders of magnitude that accompany changes in threshold voltage (i.e., from process variation or temperature changes), the optimum voltage for minimizing energy does not vary as significantly. The broad nature of the energy characteristic around the optimum point means that systems can be designed in an open loop fashion for near-optimal energy operation. However, the actual optimum V_{DD} solution for minimizing energy per operation can change over several hundred millivolts when operating parameters vary. For energy-critical systems, feedback based tracking of the optimum energy point can provide minimum energy operation even in the face of run-time changes in the energy characteristics.

In a system implementation, the dependence of delay on threshold voltage becomes a more significant issue. In traditional synchronous systems, the clock period of the entire system is set by the delay along the worst-case critical path. In a synchronous sub-threshold system, delays may vary by orders of magnitude due to global process corners and changes in temperature. Local process variation can also create more than an order of magnitude delay difference along paths with identical layout on the same die. Approaching this type of system with a "design for worst-case" mentality can decrease the overall system speed and, as a result, increase the energy at the minimum energy point. A key challenge for sub-threshold systems is to account for variations in delay without incurring the unnecessary overhead of designing for the worst-case. One approach to this problem is to use Globally Asynchronous, Locally Synchronous (GALS) designs. Breaking the system into smaller synchronous blocks lessens the impact of severe variation on the overall system.

Several open problems remain in the area of modeling the minimum energy point. First, system-level design will benefit from hierarchical application of the minimum energy point model. Different blocks in the system will exhibit different optimum voltages. The best way of grouping these blocks and/or providing separate V_{DD} regions remains an open question. A key issue in solving this problem will be the overhead required for different approaches. Perhaps the most significant circuit that must be studied in the context of overhead is the dc/dc converter. Converter design for strong-inversion designs with relatively large current requirements is a well-studied field. Relatively few works examine high efficiency dc/dc converters for low current loads and low voltages like those required for sub-threshold circuits.

One final open modeling problem involves increasing the accuracy of the minimum energy model by including higher order effects. The model that we provide gives a good estimate of the minimum energy point, but it depends on first order equations and lumped modeling parameters. Improvements in the accuracy of the model may be found by increasing its complexity. Likewise, making use of equations that span both the strong inversion and sub-threshold regions (i.e., the EKV model) will make the minimum energy model applicable to designs whose minimum energy points lie at or above the transition point between these modes of operation. Since more leaky technologies will tend to have higher optimum voltages, this modification will be necessary as scaling continues. Likewise, adapting the model to account for statistical variations at the global process level and the local transistor level will improve its accuracy.

Statistical analysis is also necessary to improve upon the sizing methodology that we produced. The sizing analysis in this thesis does not account for local transistor variations. The presence of local V_T variation will create functionality issues for digital logic at low voltages. Larger transistors will make the circuits more robust to local variations, but they will also increase both active and idle energy. The best sizing methodology in the presence of these variations is an open problem. Likewise, local variations will result in large distributions of delay through identically drawn gates across the chip. Architectures designed to adapt to these variations and to recover some of the speed lost by voltage scaling are important. Architectural advancements can also account for variation-induced errors through error detection and correction techniques. The key conclusion from sizing analysis in this thesis is that sizing is not as powerful of a knob in sub-threshold, especially in circuits with ratioed fights between currents. New circuits and architectures that can replace ratioed circuits are necessary for enabling sub-threshold operation.

Another significant issue for sub-threshold operation is system verification. Using SPICE for verifying large systems rapidly becomes infeasible when the number of process corners, temperature corners, and voltage supply values increase. Advances in CAD tools to account for this problem are necessary. These tools must also address the statistical distributions of delay and power introduced by local variations.

A UDVS proof-of-concept chip shows that UDVS is a strong candidate for tying together sub-threshold operation and higher performance operation. Future work related to UDVS should focus on system integration using this technique. Decisions related to the best interfaces among blocks operating at different effective rates and V_{DD} values will impact the system energy and delay. Selecting the best bus protocols, level converters, and dc/dc converters for a system remain open problems. Also, theoretical work related to UDVS can investigate optimum scheduling and control at the system level. The system-level analysis should consider all of the blocks and their modes of operation all the way from full shutdown to full speed active mode.

Sub-threshold memory design is fraught with challenges. The standard 6T SRAM bitcell and architecture relies on ratioed circuits for writing and compounds the higher leakage of DSM technologies by placing many leaking transistors in parallel. These characteristics make standard memories very sensitive to changes in temperature and variation in threshold voltages, and this sensitivity is amplified in the sub-threshold region. As a result, sub-threshold and low voltage operation of SRAMs requires new architectures and circuits that are more robust to variation. We have shown that sub-threshold operation is possible for SRAM in 65nm using a modified bitcell. The improvements in leakage power and active energy that result from this new cell show the advantage of going to sub-threshold operation. Although minimizing the bitcell size is a powerful motivation for SRAM, our analysis shows that larger bitcells can produce valuable reductions in power and energy consumption. Also, by increasing the overall array efficiency (i.e., by increasing the integration of cells on a bitline), the total area penalty can be reduced.

Future investigation into sub-threshold SRAM can continue to explore new circuits both at the bitcell and architecture levels. These designs should exploit advancements in technologies. For example, many DSM processes offer multiple oxide thicknesses and threshold voltages. New bitcells can utilize the advantages of such flexibility to improve operation at low voltage. Higher order concerns for sub-threshold memories, such as SER, also require investigation.

The next generation of the SRAM presented in this thesis will use better sensing techniques and improved peripheral circuits for writing to improve the yield at ultra-low voltages. Better data sensing can also speed up the read access delay. With minor changes to the peripheral circuits to improve variation immunity, we expect that the memory can function at even lower supply voltages without error.

Appendix A

Acronyms

6T	six transistor
10T	ten transistor
AOI	And/Or/Invert
BL	bitline
BPTM	Berkeley Predictive Technology Model
CDF	Cumulative Distribution Function
CMOS	Complementary MOSFET
COTS	commercial, off-the-shelf
DIBL	Drain-Induced Barrier Lowering
DIO	Data Input/Output
DRAM	Dynamic Random Access Memory
DSM	Deep Sub-Micron
DVS	Dynamic Voltage Scaling
eDRAM	embedded DRAM

EKV	Enz, Krummenacher, and Vittoz
EOFS	Environment Observation and Forecasting System
FBB	Forward Body Bias
FET	Field Effect Transistor
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FS	Fast nMOS, Slow pMOS
GALS	Globally Asynchronous, Locally Synchronous
GIDL	Gate-Induced Drain Leakage
IC	Integrated Circuit
iid	independent, identically distributed
LFSR	Linear Feedback Shift Register
LVD	Local Voltage Dithering
M-C	Monte-Carlo
MEMS	micro electromechanical systems
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor
MTCMOS	multi-threshold CMOS
nMOS	n-type MOSFET
PDA	Personal Digital Assistant
PDF	Probability Density Function
pMOS	p-type MOSFET

RBB	Reverse Body Bias
RBL	read bitline
RWL	read wordline
SER	Soft Error Rate
SF	Slow nMOS, Fast pMOS
SHM	Structural Health Monitoring
SNM	Static Noise Margin
SOI	Silicon on Insulator
SRAM	Static Random Access Memory
SS	Strong nMOS, Strong pMOS
STC	stacked capacitor
SW	Strong nMOS, Weak pMOS
TT	Typical nMOS, Typical pMOS
UDVS	Ultra-Dynamic Voltage Scaling
VCO	voltage-controlled oscillator
VTC	Voltage Transfer Characteristic
WINS	Wireless Integrated Network Sensors
WL	wordline
WS	Weak nMOS, Strong pMOS
WW	Weak nMOS, Weak pMOS

Appendix B

Lambert W Function

Equation (B.1) shows the Lambert W Function.

$$W = \mathrm{lambertW}(x) \tag{B.1}$$

The Lambert W Function gives the solution to the equation:

$$We^{W} = x \tag{B.2}$$

in the same way that $W = \ln x$ is the solution to $e^{W} = x$. Figure B-1 shows $e^{W} = x$ and its inverse, $W = \ln x$, for reference.

Figure B-2 plots equations (B.1) and (B.2). For real $x \geq 0$, (B.1) has exactly one real solution. For real $-e^{-1} < x < 0$, there are exactly two real solutions, called branches. The upper branch increases monotonically in $[-1, \infty)$ for $x \in [-e^{-1}, \infty)$, and the lower branch decreases monotonically in $[-1, -\infty)$ for $x \in [-e^{-1}, 0)$ [147].

For the solution in Equation (2.10), the argument to lambertW is always negative, so two real solutions exist subject to the constraint $\beta > -e^{-1}$. The lower branch gives the minimum energy solution, and the upper branch solution is the local maximum.

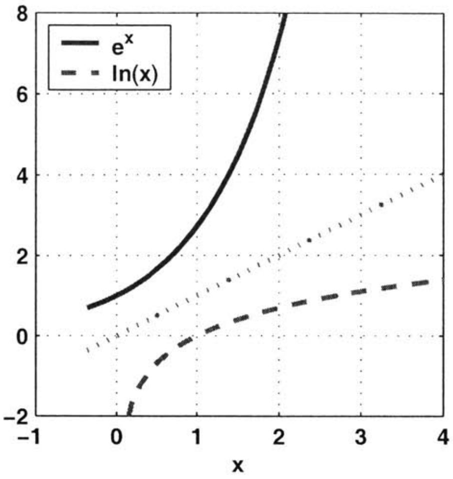


Figure B-1: The natural log function, $W = \ln x$, provides the inverse of $e^W = x$ and is analogous to the Lambert W Function.

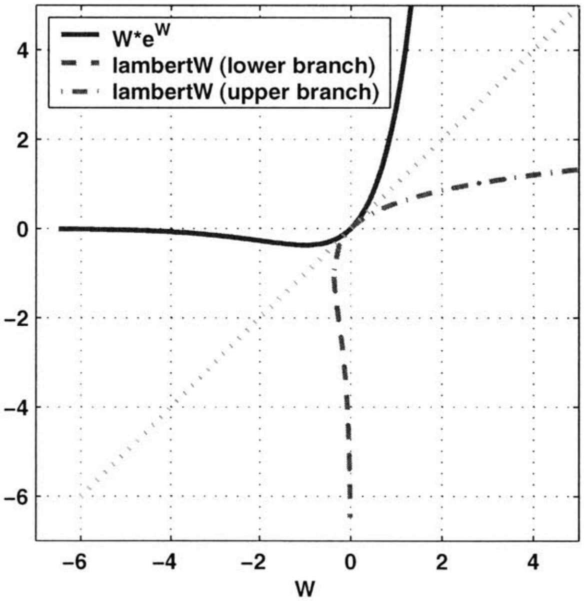


Figure B-2: The Lambert W Function, $W = \text{lambertW}(x)$, gives the solution to $We^W = x$. This plot shows the inverse relationship between these two equations, and it shows the upper and lower branch solutions for the Lambert W Function.

Bibliography

[1] [Online]. Available: http://www.eas.asu.edu/~ptm

[2] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design," in *Custom Integrated Circuits Conference (CICC) Digest of Technical Papers*, Oct. 2000, pp. 201–204.

[3] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for Low-Power and High-Speed Applications," in *ACM/IEEE Design Automation Conference (DAC) Digest of Technical Papers*, 2000, pp. 469–474.

[4] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Sub-threshold Circuit Techniques," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 2004, pp. 292–293.

[5] V. Gutnik and A. Chandrakasan, "Embedded Power Supply for Low-Power DSP," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 5, no. 4, pp. 425–435, Dec. 1997.

[6] E. Seevinck, F. 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.

[7] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa, and K. Osadal, "A 300MHz 25μA/Mb Leakage On-Chip SRAM Module Featuring Process-Variation Immunity and Low-Leakage-Active Mode for Mobile-Phone Application Processor," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 2004, pp. 494–495.

[8] A. Bhavnagarwala, S. Kosonocky, S. Kowalczyk, R. Joshi, Y. Chan, U. Srinivasan, and J. Wadhwa, "A Transregional CMOS SRAM with Single, Logic V_{DD} and Dynamic Power Rails," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2004, pp. 292–293.

[9] K. Noda, K. Matsui, K. Imai, K. Inoue, K. Tokashiki, H. Kawamoto, K. Yoshida, K. Takeda, N. Nakamura, T. Kimura, H. Toyoshima, Y. Koishikawa, S. Maruyama, T. Saitoh, and T. Tanigawa, "A 1.9-μm^2 Loadless CMOS Four-Transistor SRAM Cell in a 0.18-μm Logic Technology," in *International Electron Devices Meeting (IEDM) Digest of Technical Papers*, 1998, pp. 643–646.

[10] T.-H. Joubert, E. Seevinck, and M. du Plessis, "A CMOS Reduced-Area SRAM Cell," in *International Symposium on Circuits and Systems (ISCAS) Digest of Technical Papers*, 2000, pp. III335–III338.

[11] H. Tran, "Demonstration of 5T SRAM and 6T Dual-Port RAM Cell Arrays," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1996, pp. 68–69.

[12] N. Azizi and F. N. Najm, "An Asymmetric SRAM Cell to Lower Gate Leakage," in *International Symposium on Quality Electronic Design (ISQED) Digest of Technical Papers*, 2004, pp. 534–539.

[13] G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[14] ——, "No Exponential is Forever: But 'Forever' Can Be Delayed!" in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2003, pp. 20–23.

[15] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance Microprocessors," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, 1999, pp. 163–168.

[16] S. Borkar, "Low-Voltage Design for Portable Systems - Leakage Reduction in Digital CMOS Circuits," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2002, pp. 577–580.

[17] J. Burr and A. Peterson, "Ultra Low Power CMOS Technology," in *3rd NASA Symposium on VLSI Design*, 1991, pp. 4.2.1–4.2.13.

[18] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal Supply and Threshold Scaling for Sub-threshold CMOS Circuits," in *IEEE Computer Society Annual Symposium on VLSI*, Apr. 2002, pp. 7–11.

[19] G. Asada, M. Dong, T. Lin, F. newberg, G. Pottie, W. Kaiser, and H. Marcy, "Wireless Integrated Network Sensors: Low Power Systems on a Chip," in *European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, 1998, pp. 9–16.

[20] A. Cerpa, J. Elson, D. Estrin, L. Girod, M. Hamilton, and J. Zhao, "Habitat Monitoring: Application Driver for Wireless Communications Technology," in *Proceedings of the ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean*, 2001, pp. 20–41.

[21] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," in *ACM International Workshop on Wireless Sensor Networks and Applications (WSNA)*, Sept. 2002, pp. 88–97.

[22] E. Biagioni and K. Bridges, "The Application of Remote Sensor Technoloy to Assist the Recovery of Rare and Endangered Species," *Special Issue on Distributed Sensor Networks for the International Journal of High Performance Computing Applications*, vol. 16, no. 3, pp. 315–324, Aug. 2002.

[23] [Online]. Available: http://www.ccalmr.ogi.edu/CORIE/

[24] [Online]. Available: http://www.alertsystems.org/

[25] L. Schwiebert, S. Gupta, and J. Weinmann, "Research Challenges in Wireless Networks of Biomedical Sensors," in *Mobile Computing and Networking*, 2001, pp. 151–165.

[26] K. Chintalapudi, E. Johnson, and R. Govindan, "Structural Damage Detection Using Wireless Sensor-Actuator Networks," in *Proceedings of the IEEE International Symposium on Intelligent Control*, June 2005, pp. 322–327.

[27] R. Measures, "Fiber Optic Structural Monitoring of Bridges," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, 1997, pp. 600–602.

[28] P. Foote and I. Read, "Optical Sensors for Aerospace Structural Monitoring," in *IEE Colloquium on Optical Techniques for Structural Monitoring*, Apr. 1995, pp. 2/1–2/6.

[29] A. Basharat, N. Catbas, and M. Shah, "A Framework for Intelligent Sensor Network with Video Camera for Structural Health Monitoring of Bridges," in *Proceedings of the IEEE Conference on Pervasive Computing and Communications (PerCom)*, Mar. 2005, pp. 385–389.

[30] N. M. Su, H. Park, E. Bostrom, J. Burke, M. B. Srivastava, and D. Estrin, "Augmenting Film and Video Footage with Sensor Data," in *IEEE International Conference on Pervasive Computing and Communications*, 2004, pp. 3–12.

[31] N. Xu, "A Survey of Sensor Network Applications," University of Southern California, Tech. Rep., 2003. [Online]. Available: http://enl.usc.edu/ ningxu/papers/survey.pdf

[32] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, "Next Century Challenges: Scalable Coordination in Sensor Networks," in *Proceedings of the ACM MobiCom*, 1999, pp. 263–270.

[33] M. Hempstead, N. Tripathi, P. Mauro, G. Wei, and D. Brooks, "An Ultra Low Power System Architecture for Sensor Network Applications," in *International Symposium on Computer Architecture (ISCA) Digest of Technical Papers*, 2005, pp. 208–219.

[34] L. Nazhandali, B. Zhai, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, T. Austin, and D. Blaauw, "Energy Optimization of Subthreshold-Voltage Sensor Network Processors," in *International Symposium on Computer Architecture (ISCA) Digest of Technical Papers*, 2005, pp. 197–207.

[35] [Online]. Available: http://www.xbow.com

[36] J. Polastre, R.Szewczyk, and D. Culler, "Telos: Enabling Ultra-Low Power Wireless Research," in *Proceedings of the International Symposium on Information Processing in Sensor Networks*, Apr. 2005, pp. 364–369.

[37] J. Kahn, R. Katz, and K. Pister, "Next Century Challenges: Mobile Networking for Smart Dust," in *Proceedings of the ACM MobiCom*, 1999, pp. 271–278.

[38] R. Min, M. Bhardwaj, S.-H. Cho, A. Sinha, E. Shih, A. Wang, and A. Chandrakasan, "An Architecture for a Power-Aware Distributed Microsensor Node," in *Proceedings of the IEEE Workshop on Signal Processing Systems (SiPS)*, Oct. 2000, pp. 581–590.

[39] S. Roundy, P. Wright, and J. Rabaey, "A Study of Low Level Vibrations as a Power Source for Wireless Sensor Nodes," *Computer Communications*, vol. 26, no. 11, pp. 1131–1144, July 2003.

[40] H. Kulah and K. Najafi, "An Electromagnetic Micro Power Generator for Low-Frequency Environmental Vibrations," in *Proceedings of the IEEE International Conference for Micro Electro Mechanical Systems (MEMS)*, Jan. 2004, pp. 237–240.

[41] S. Meninger, J. Mur-Miranda, R. Amirtharajah, A. Chandrakasan, and J. Lang, "Vibration-to-electric Energy Conversion," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 1, pp. 64–76, Feb. 2001.

[42] H. Bottner, J. Nurnus, A. Gavrikov, G. Kuhner, M. Jagle, C. Kunzel, D. Eberhard, G. Plescher, A. Schubert, and K.-H. Schlereth, "New Thermoelectric Components Using Microsystems Technologies," *IEEE/ASME Journal of Microelectromechanical Systems*, vol. 13, no. 3, pp. 414–420, June 2004.

[43] *Panasonic Solar Cells Technical Handbook '98/'99*, Matsushita Battery Industrial Co., Ltd., Aug. 1998.

[44] B. H. Calhoun, D. Daly, N. Verma, D. Finchelstein, D. D. Wentzloff, A. Wang, S.-H. Cho, and A. Chandrakasan, "Design Considerations for Ultra-Low Energy Wireless Microsensor Nodes," vol. 54, no. 6, pp. 727–740, June 2005.

[45] 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 International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2000, pp. 362–363.

[46] R. M. Swanson and J. D. Meindl, "Ion-Implanted Complementary MOS Transistors in Low-Voltage Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-7, no. 2, pp. 146–153, Apr. 1972.

[47] R. Swanson, "Complementary MOS Transistors in Micropower Circuits," Ph.D. dissertation, Stanford University, 1974.

[48] E. Vittoz and J. Fellrath, "CMOS Analog Integrated Circuits Based on Weak-Inversion Operation," *IEEE Journal of Solid-State Circuits*, vol. SC-12, pp. 224–231, June 1977.

[49] C. Mead, *Analog VLSI and Neural Systems.* Addison-Wesley, 1989.

[50] R. Daniels and R. Burgess, "The Electronic Wristwatch: an Application for Si-Gate CMOS ICs," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 1971, pp. 62–63.

[51] E. Vittoz, B. Gerber, and F. Leuenberger, "Silicon-Gate CMOS Frequency Divider for the Electronic Wrist Watch," vol. SC-7, pp. 100–104, 1972.

[52] J. Burr and A. Peterson, "Energy Considerations in Multichip-Module Based Multiprocessors," in *IEEE International Conference on Computer Design (ICCD) Digest of Technical Papers*, 1991, pp. 593–600.

[53] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992.

[54] D. Liu and C. Svensson, "Trading Speed for Low Power by Choice of Supply and Threshold Voltages," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 1, pp. 10–17, Jan. 1993.

[55] J. Burr, "Cryogenic Ultra Low Power CMOS," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, 1995, pp. 82–83.

[56] E. Vittoz, "Low-Power Design: Ways to Approach the Limits," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 1994, pp. 14–18.

[57] G. Schrom and S. Selberherr, "Ultra-Low-Power CMOS Technologies," in *International Semiconductor Conference (CAS) Digest of Technical Papers*, 1996, pp. 237–246.

[58] H. Soeleman and K. Roy, "Ultra-Low Power Digital Subthreshold Logic Circuits," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, 1999, pp. 94–96.

[59] H. Soeleman, K. Roy, and B. Paul, "Sub-Domino Logic: Ultra-Low Power Dynamic Sub-Threshold Digital Logic," in *International Conference on VLSI Design (VLSI-Design) Digest of Technical Papers*, Jan. 2001, pp. 211–214.

[60] H. Kim and K. Roy, "Ultra-Low Power DLMS Adaptive Filter for Hearing Aid Applications," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, Aug. 2001, pp. 352–357.

[61] C. H. Kim, H. Soeleman, and K. Roy, "Ultra-Low-Power DLMS Adaptive Filter for Hearing Aid Applications," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 4, pp. 716–730, Aug. 2003.

[62] A. Bryant, J. Brown, P. Cottrell, M. Ketchen, J. Ellis-Monaghan, and J. Nowak, "Low-Power CMOS at Vdd=4kT/q," in *Device Research Conference*, June 2001, pp. 22–23.

[63] M. Deen, H. Kazemeini, and S. Naseh, "Ultra-low Power VCOs - Performance Characteristics and Modeling," in *IEEE Internationl Caracas Conference on Devices, Circuits and Systems Digest of Technical Papers*, Apr. 2002, pp. C033-1 to C033–8.

[64] C. Enz, F. Krummenacher, and E. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications," *Journal on Analog Integrated Circuits and Signal Processing*, pp. 83–114, July 1995.

[65] V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar, "Techniques for Leakage Power Reduction," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. Bowhill, and F. Fox, Eds. IEEE Press, 2001, ch. 3, pp. 46–62.

[66] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[67] E. Vittoz, "Weak Inversion for Ultimate Low-Power Logic," in *Low-Power Electronics Design*, C. Piguet, Ed. CRC Press, 2005, pp. 16–1 to 16–18.

[68] R. Gonzalez, B. Gordon, and M. Horowitz, "Supply and Threshold Voltage Scaling for Low Power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, Aug. 1997.

[69] M. Stan, "Optimal Voltages and Sizing for Low Power," in *International Conference on VLSI Design (VLSI-Design) Digest of Technical Papers*, 1999, pp. 428–433.

[70] J. Kao, M. Mayazaki, and A. Chandrakasan, "A 175-mV Multiply-Accumulate Unit Using an Adaptive Supply Voltage and Body Bias Architecture," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1545–1554, Nov. 2002.

[71] A. Bhavnagarwala, B. Austin, K. Bowman, and J. Meindl, "A Minimum Total Power Methodology for Projecting Limits on CMOS GSI," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 235–251, June 2000.

[72] R. Brodersen, M. Horowitz, D. Markovic, B. Nikolic, and V. Stojanovic, "Methods for True Power Minimization," in *IEEE International Conference on Computer-Aided Design (ICCAD) Digest of Technical Papers*, 2002, pp. 35–42.

[73] M. Miura-Mattausch, M. Suetake, J. Mattausch, S. Kumashiro, and N. Shigyo, "Physical Modeling of the Reverse-Short-Channel Effect for Circuit Simulation," *IEEE Transactions on Electron Devices*, vol. 48, no. 10, pp. 2449–2452, Oct. 2001.

[74] A. Ono, R. Ueno, and I. Sakai, "TED Control Technology for Suppression of Reverse Narrow Channel Effect in 0.1μm MOS Devices," in *International Electron Devices Meeting (IEDM) Digest of Technical Papers*, 1997, pp. 227–230.

[75] B. H. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, 2004, pp. 90–95.

[76] B. 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. 9, pp. 1778–1786, Sept. 2005.

[77] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," in *ACM/IEEE Design Automation Conference (DAC) Digest of Technical Papers*, 2004, pp. 868–873.

[78] A. Bellaouar, A. Fridi, M. Elmasry, and K. Itoh, "Supply Voltage Scaling for Temperature Insensitive CMOS Circuit Operation," *IEEE Transactions on Circuits and Systems*, vol. 45, no. 3, pp. 415–417, Mar. 1998.

[79] C. Piguet, "Design of Low-Power Libraries," in *International Conference on Electronics, Circuits and Systems (ICECS) Digest of Technical Papers*, vol. 2, Sept. 1998, pp. 175–180.

[80] C. Piguet, J.-M. Masgonty, S. Cserveny, C. Arm, and P.-D. Pfister, "Low-Power Low-Voltage Library Cells and Memories," in *International Conference on Electronics, Circuits and Systems (ICECS) Digest of Technical Papers*, 2001, pp. 1521–1524.

[81] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Device Sizing for Minimum Energy Operation in Subthreshold Circuits," in *Custom Integrated Circuits Conference (CICC) Digest of Technical Papers*, Oct. 2004, pp. 95–98.

[82] P. Macken, M. Degrauwe, M. V. Paemel, and H. Oguey, "A Voltage Reduction Technique for Digital Systems," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 1990, pp. 238–239.

[83] [Online]. Available: http://www.intel.com/design/intelxscale/

[84] K. Nowka, G. Carpenter, E. MacDonald, H. Ngo, B. Brock, K. Ishii, T. Nguyen, and J. Burns, "A 0.9V to 1.95V Dynamic Voltage-Scalable and Frequency-Scalable 32b PowerPC Processor," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2002, pp. 340–341.

[85] [Online]. Available: http://www.transmeta.com/crusoe/index.html

[86] H. Kawaguchi, G. Zhang, S. Lee, Y. Shin, and T. Sakurai, "A Controller LSI for Realizing VDD-Hopping Scheme with Off-the-Shelf Processors and Its Application to MPEG4 System," *IEICE Transactions on Electronics*, vol. E85-C, no. 2, pp. 263–271, Feb. 2002.

[87] B. Calhoun and A. Chandrakasan, "Ultra-Dynamic Voltage Scaling (UDVS) Using Sub-threshold Operation and Local Voltage Dithering in 90nm CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 300–301.

[88] A. Chandrakasan, V. Gutnik, and T. Xanthopoulos, "Data Driven Signal Processing: An Approach for Energy Efficient Computing," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, 1996, pp. 347–352.

[89] L. Chandrasena, P. Chandrasena, and M. Liebelt, "An Energy Efficient Rate Selection Algorithm for Voltage Quantized Dynamic Voltage Scaling," in *International Symposium on System Synthesis*, Oct. 2001, pp. 124–129.

[90] S. Lee and T. Sakurai, "Run-time Power Control Scheme Using Software Feedback Loop for Low-Power Real-time Applications," in *Asia and South-Pacific Design Automation Conference*, Jan. 2000, pp. 381–386.

[91] ——, "Run-time Voltage Hopping for Low-power Real-time Systems," in *ACM/IEEE Design Automation Conference (DAC) Digest of Technical Papers*, June 2000, pp. 806–809.

[92] H. Kawaguchi, K. Kanda, K. Nose, S. Hattori, D. Dwi, D. Antono, D. Yamada, T. Miyazaki, K. Inagaki, T. Hiramoto, and T. Sakurai, "A 0.5V, 400MHz, VDD-Hopping Processor with Zero-VTH FD-SOI Technology," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2003, pp. 106–107.

[93] L. Chandrasena and M. Liebelt, "Energy Minimization in Dynamic Supply Voltage Scaling Systems Using Data Dependent Voltage Level Selection," in

International Symposium on Circuits and Systems (ISCAS) Digest of Technical Papers, 2000, pp. 525–528.

[94] N. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 2, pp. 167–184, Feb. 2004.

[95] B. H. Calhoun and A. Chandrakasan, "Standby Power Reduction Using Dynamic Voltage Scaling and Canary Flip-Flop Structures," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1504–1511, Sept. 2004.

[96] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," in *International Symposium on Quality Electronic Design (ISQED) Digest of Technical Papers*, 2004, pp. 55–60.

[97] K. Itoh, "Low-Voltage Memories for Power-Aware Systems," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, Aug., 2002, pp. 1–6.

[98] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Prentice Hall, 2003.

[99] K. Kanda, T. Miyazaki, M. K. Sik, H. Kawaguchi, and T. Sakurai, "Two Orders of Magnitude Leakage Power Reduction of Low Voltage SRAM's by Row-by-Row Dynamic V_{DD} Control (RRDV) Scheme," in *IEEE International ASIC/SOC Conference*, Sept. 2002, pp. 381–385.

[100] T. Enomoto, Y. Oka, and H. Shikano, "A Self-Controllable Voltage Level (SVL) Circuit and its Low-Power High-Speed CMOS Circuit Applications," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 7, pp. 1220–1226, July 2003.

[101] M. Yamaoka, K. Osada, and K. Ishibashi, "0.4-V Logic Library Friendly SRAM Array Using Rectangular-Diffusion Cell and Delta-Boosted-Array-Voltage Scheme," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2002, pp. 170–173.

[102] ——, "0.4-V Logic-Library-Friendly SRAM Array Using Rectangular-Diffusion Cell and Delta-Boosted-Array Voltage Scheme," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 6, pp. 934–940, June 2004.

[103] A. Bhavnagarwala, A. Kapoor, and J. Meindl, "Dynamic-Threshold CMOS SRAM Cells for Fast, Portable Applications," in *IEEE International ASIC/SOC Conference*, Sept. 2000, pp. 359–363.

[104] K. Itoh, A. Fridi, A. Bellaouar, and M. Elmasry, "A Deep Sub-V, Single Power-Supply SRAM Cell with Multi-V_T, Boosted Storage Node and Dynamic Load,"

in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1996, pp. 132–133.

[105] H. Yamauchi, T. Iwata, H. Akamatsu, and A. Matsuzawa, "A 0.8V/100MHz/sub-5mW-operated Mega-Bit SRAM Cell Architecture with Charge-Recycle Offset-Source Driving (OSD) Scheme," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1996, pp. 126–127.

[106] 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.

[107] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Yang, B. Zheng, and M. Bohr, "A SRAM Design on 65nm CMOS Technology with Integrated Leakage Scheme," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2004, pp. 294–295.

[108] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, and H. Makino, "A 90nm Dual-Port SRAM with 2.04μm^2 8T-Thin Cell Using Dynamically-Controlled Column Bias Scheme," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 2004, pp. 508–509.

[109] A. Agarwal, H. Li, and K. Roy, "A Single-V_t Low-Leakage Gated-Ground Cache for Deep Submicron," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, Feb. 2003.

[110] K. Kanda, H. Sadaaki, and T. Sakurai, "90% Write Power-Saving SRAM Using Sense-Amplifying Memory Cell," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 6, pp. 927–933, June 2004.

[111] K. Osada, J. L. Shin, M. Khan, Y. Liou, K. Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi, "Universal-Vdd 0.65–2.0-V 32-kB Cache Using a Voltage-Adapted Timing-Generation Scheme and a Lithographically Symmetrical Cell," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 1738–1744, Nov. 2001.

[112] S. Ikeda, Y. Yoshida, K. Ishibashi, and Y. Mitsui, "Failure Analysis of 6T SRAM on Low-Voltage and High-Frequency Operation," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1270–1276, May 2003.

[113] C. Lage, D. Burnett, T. McNelly, K. Baker, A. Bormann, D. Dreier, and V. Soorholtz, "Soft Error Rate and Stored Charge Requirements in Advanced High-Density SRAMs," in *International Electron Devices Meeting (IEDM) Digest of Technical Papers*, 1993, pp. 821–824.

[114] K. Osada, K. Yamaguchi, Y. Saitoh, and T. Kawahara, "SRAM Immunity to Cosmic-Ray-Induced Multierrors Based on Analysis of an Induced Parasitic Bipolar Effect," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 5, pp. 827–833, May 2004.

[115] H. Kawaguchi, Y. Itaka, and T. Sakurai, "Dynamic Cut-off Scheme for Low-Voltage SRAM's," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1998, pp. 140–141.

[116] M. Yamaoka, K. Osada, R. Tsuchiya, M. Horiuchi, S. Kimura, and T. Kawahara, "Low Power SRAM Menu for SOC Application Using Yin-Yang-Feedback Memory Cell Technology," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2004, pp. 288–291.

[117] K. Zhang, K. Hose, V. De, and B. Senyk, "The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18μm," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2000, pp. 226–227.

[118] F. Hamzaoglu, Y. Ye, A. Keshavarzi, K. Zhang, S. Narendra, S. Borkar, M. Stan, and V. De, "Analysis of Dual-V_T SRAM Cells with Full-Swing Single-Ended Bit Line Sensing for On-Chip Cache," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 10, no. 2, pp. 91–95, Apr. 2002.

[119] J. Masgonty, S. Cserveny, and S. Piguet, "Low-Power SRAM and ROM Memories," in *International Workshop on Power and Timing, Modeling, Optimization, and Simulation (PATMOS) Digest of Technical Papers*, 2001, pp. 7.4.1–7.4.7.

[120] S. Cserveny, J.-M. Masgonty, and C. Piguet, "Noise Margin in Low Power SRAM Cells," in *International Workshop on Power and Timing, Modeling, Optimization, and Simulation (PATMOS) Digest of Technical Papers*, Sept. 2004, pp. 889–898.

[121] M. Morimura and N. Shibata, "A 1-V 1-Mb SRAM for Portable Equipment," in *International Symposium on Low-Power Electronics and Design (ISLPED) Digest of Technical Papers*, 1996, pp. 61–66.

[122] K. Ishibashi, K. Takasugi, T. Yamanaka, T. Hashimoto, and K. Sasaki, "A 1-V TFT-load SRAM using a Two-Step Word-Voltage Method," *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 1519–1524, 1992.

[123] K. Agawa, H. Hara, T. Takayanagi, and T. Kuroda, "A Bitline Leakage Compensation Scheme for Low-Voltage SRAMs," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 5, pp. 726–734, May 2001.

[124] A. Alvandpour, D. Somasekhar, R. Krishnamurthy, V. De, S. Borkar, and C. Svensson, "Bitline Leakage Equalization for Sub-100nm Caches," in *European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, 2003, pp. 401–404.

[125] B. Amrutur, "Fast Low Power SRAMs," Ph.D. dissertation, Stanford University, Computer Systems Laboratory, Stanford, CA, 1999.

[126] J. Ku, S. Siu, M. Yazdani, Y. Lih, W. Lu, and A. Desroches, "A 2.25 Gbytes/s 1Mbit Smart Cache SRAM," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1995, pp. 17–18.

[127] A. Bhavnagarwala, S. Kosonocky, and J. Meindl, "Interconnect-Centric Array Architectures for Minimum SRAM Access Time," in *IEEE International Conference on Computer Design (ICCD) Digest of Technical Papers*, Sept. 2001, pp. 400–405.

[128] B. S. Amrutur and M. A. Horowitz, "Fast low-power decoders for RAMs," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 1506–1515, Oct. 2001.

[129] T. Uetake, Y. Maki, T. Nakadai, K. Yoshida, M. Susuki, and R. Nanjo, "A 1.0ns access 770MHz 36Kb SRAM Macro," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1999, pp. 109–110.

[130] A. R. Pelella, P. Lu, Y. H. Chan, W. V. Huott, U. Bakhru, S. Kowalczyk, P. Patel, J. Rawlins, and P. T. Wu, "A 2ns access, 500MHz 288Kb SRAM MACRO," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1996, pp. 128–129.

[131] S. D. Beer, M. du Plessis, and E. Seevinck, "An SRAM Array Based on a Four-Transistor CMOS SRAM Cell," *IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications*, vol. 50, no. 9, pp. 1203–1208, Sept. 2003.

[132] I. Carlson, S. Andersson, S. Natarajan, and A. Alvandpour, "A High Density, Low Leakage, 5T SRAM for Embedded Caches," in *European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, Sept. 2004, pp. 215–218.

[133] K. Itoh, T. Watanabe, S. Kimura, and T. Sakata, "Reviews and Prospects of High-Density RAM Technology," in *International Semiconductor Conference (CAS) Digest of Technical Papers*, Oct. 2000, pp. 13–22.

[134] S. Iyer and H. Kalter, "Embedded DRAM Technology: System-Level Designs that Include DRAM and Logic on the Same IC Pay Off in Higher Memory Bandwidth and Superior Performance," *IEEE Spectrum*, vol. 36, no. 4, pp. 56–64, Apr. 1999.

[135] S. Crowder, R. Hannon, H. Ho, D. Sinitsky, S. Wu, K. Winstel, B. Khan, S. Stiffler, and S. Iyer, "Integration of Trench DRAM into a High-Performance $0.18/mum$ Logic Technology with Copper BEOL," in *International Electron Devices Meeting (IEDM) Digest of Technical Papers*, 1998, pp. 1017–1020.

[136] K. Hardee, F. Jones, D. Butler, M. Parris, M. Mound, H. Calendar, G. Jones, L. Aldrich, C. Gruenschlaeger, M. Miyabayashi, K. Taniguchi, and T. Arakawa, "A 0.6V 205MHz 19.5ns tRC 16Mb Embedded DRAM," in *IEEE International*

Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, Feb. 2004, pp. 200–201.

[137] D. Somasekhar, S. Lu, B. Bloechel, G. Dermer, K. Lai, S. Borkhar, and V. De, "A 10Mbit, 15GBytes/sec Bandwidth 1T DRAM Chip with Planar MOS Storage Capacitor in an Unmodified 150nm Logic Process for High-Density On-Chip Memory Applications," in *European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, Sept. 2005, pp. 355–358.

[138] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "Low-Power Embedded SRAM Modules with Expanded Margins for Writing," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 480–481.

[139] 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, Apr. 2001.

[140] B. Calhoun and A. Chandrakasan, "Analyzing Static Noise Margin for Subthreshold SRAM in 65nm CMOS," in *European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, Sept. 2005, pp. 363–366.

[141] B. Cheng, S. Roy, and A. Asenov, "The Impact of Random Doping Effects on CMOS SRAM Cell," in *European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, 2004, pp. 219–222.

[142] R. Keyes, "The Effect of Randomness in the Distribution of Impurity Atoms on FET Threshold," *Applied Physics A: Materials Science and Processing*, vol. 8, pp. 251–259, 1975.

[143] S. Mukhopadhyay, H. Mahmoodi-Meimand, and K. Roy, "Modeling and Estimation of Failure Probability due to Parameter Variations in Nano-Scale SRAMs for Yield Enhancement," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2004, pp. 64–67.

[144] K. Takeuchi, R. Koh, and T. Mogami, "A Study of Threshold Voltage Variation for Ultra-Small Bulk and SOI CMOS," *IEEE Transactions on Electron Devices*, vol. 48, no. 9, pp. 1995–2001, Sept. 2001.

[145] 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-Vdd and High-Speed Applications," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 478–479.

[146] I. Kim, Y. Zorian, G. Komoriya, H. Pham, F. Higgins, and J. Lewandowski, "Built in Self Repair for Embedded High Density SRAM," in *International Test Conference (ITC) Digest of Technical Papers*, 1998, pp. 1112–1119.

[147] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert W Function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.