Power Management and SRAM for Energy-Autonomous and Low-Power Systems

by

Gregory K. Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2011

Doctoral Committee:

Professor Dennis M. Sylvester, Chair
Professor David Blaauw
Associate Professor Jerome P. Lynch
Assistant Professor David D. Wentzloff
Senior Principal Engineer Ram K. Krishnamurthy, Intel Corporation

TABLE OF CONTENTS

CHAPTERS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER 1

Introduction

During the evolution of the IC industry, a new class of computers has emerged approximately every decade. This trend was predicted by Bell's Law and has accurately traced the development of mainframe computers in the 1960s, to personal computers in the 1980s, to mobile computing devices in the 2000s (Fig. 1.1) [1]. With each computer class, size reduces by roughly 100x, lower cost makes computers more numerous and accessible, and IC developments increase processing power. For example, a company may only have one mainframe that occupies a whole room, but each of its employees has a smartphone that easily fits inside their pocket. Bell's Law points to the development of even smaller devices, namely millimeter-scale sensor motes.

Pervasive sensors will provide computing resources with a new, unprecedented level of accessibility. Today, we interface directly with our laptops and smartphones, issuing direct commands that the computers blindly follow. Miniature ubiquitous computers could one day enable cognitive awareness in wireless networks. These future devices will autonomously monitor environmental conditions and report unsolicited but useful information. In this way, pervasive sensors provide us with 6^{th}, 7^{th}, and 8^{th} senses beyond our biological sight, sound, smell, taste and feel. They allow us to "feel" corrosion in our surrounding infrastructure and imperceptible seismic vibrations that might damage it. These senses make it easier to detect and correct weaknesses in bridges and buildings before catastrophic failure. The microsensors improve on our innate senses, giving us the ability to detect single degree changes in temperature and single percent changes in humidity. This accuracy enables smart buildings, where heating, ventilation and air conditioning (HVAC) systems provide more-comfortable living and working conditions with lower energy costs. Ubiqui-

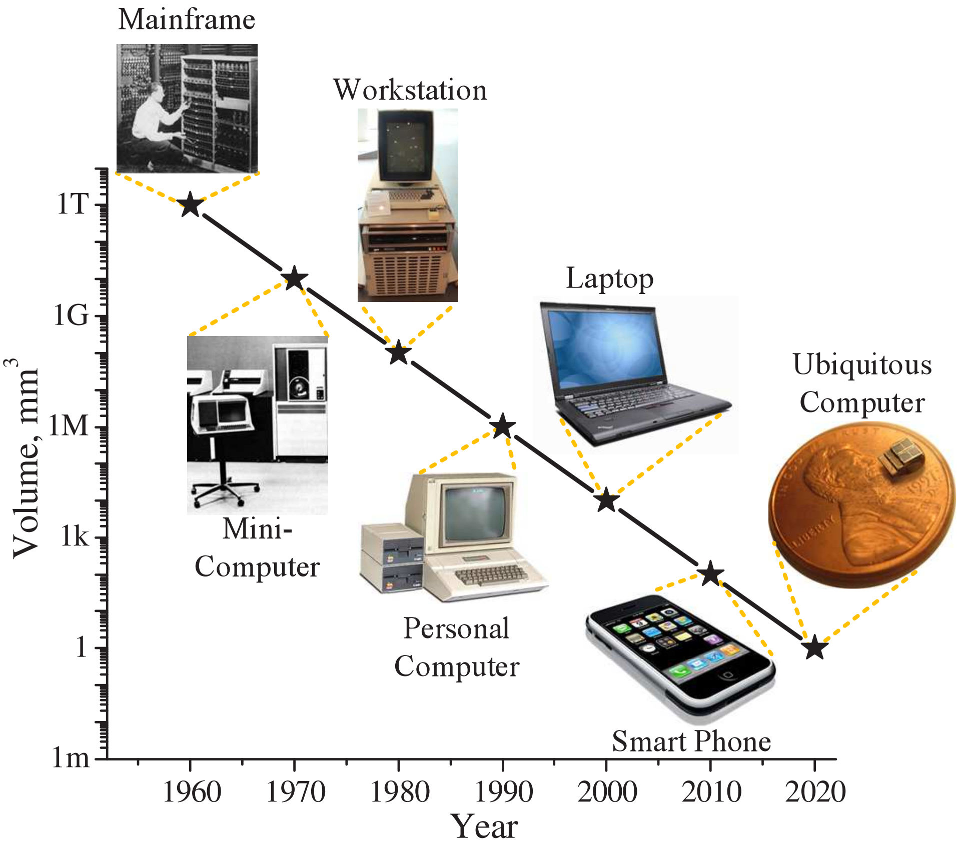


Figure 1.1: Bell's Law predicts a new class of smaller, more numerous computer systems every decade. Massively deployed cubic-millimeter microsensors such as the presented IOP monitor are the next step in this evolution of the IC industry.

tous computers give us clairvoyance by extending our biological senses beyond their normal reach. They allow us to detect shock waves when our package is dropped halfway across the world, so we can ensure that the products are not damaged during transport.

Ubiquitous sensors must be inexpensive and seamlessly integrated with our surroundings to be practical. Today's sensors are composed of milliWatt components, including commercial-off-the-shelf processors, analog components, and discrete passives, assembled on a printed circuit board. These components and their assembly all have costs, decreasing the economic viability of massively deployed sensor networks. Miniaturized densely integrated microsensors leverage developments in integrated circuit processing technology to combine many of these components to reduce cost. In general, large sensors are obtrusive and aesthetically displeasing, but for many applications, millimeter-scale size is vital. For example, thinner surveillance motes more easily avoid detection. Smaller medical implants make implantation less invasive and reduce trauma to the surrounding tissue.

Microsystem components, including power sources, sensing elements and integrated circuits, can be miniaturized, but additional system-level power constraints emerge. Because of today's devices' high power consumption, they are powered by bulky power sources such as AA batteries to obtain a lifetime of months or years. Energy harvesting sensors collect energy from their environment and can achieve energy-autonomous operation, where the device harvests more energy that it uses. However large harvesting elements, such as centimeter-scale solar cells, are needed to meet today's sensors' power demands. Millimeter-scale power sources can hold very little energy on-sensor. For example, a 12mm^2 thin-film Li battery only has an energy capacity of 12μAh [2], in contrast to a 3000mAh AA alkaline battery [3]. In addition, millimeter-scale energy harvesters deliver only a few microWatts of power with sporadic availability. These constraints limit the average power consumption of millimeter-scale sensors to nanoWatts or picoWatts (Fig. 1.2). A microsystem achieving these load powers requires a new class of low-power electronics that use design techniques such as subthreshold operation, weak-inversion biasing, aggressive power gating, and efficient low-load power delivery [4].

A canonical sensor system reads the digitized output of sensing elements (Fig. 1.3). A microprocessor (μP) performs digital signal processing and compresses the raw sensor data before logging its result into memory, such as SRAM. The end user accesses the data from memory using

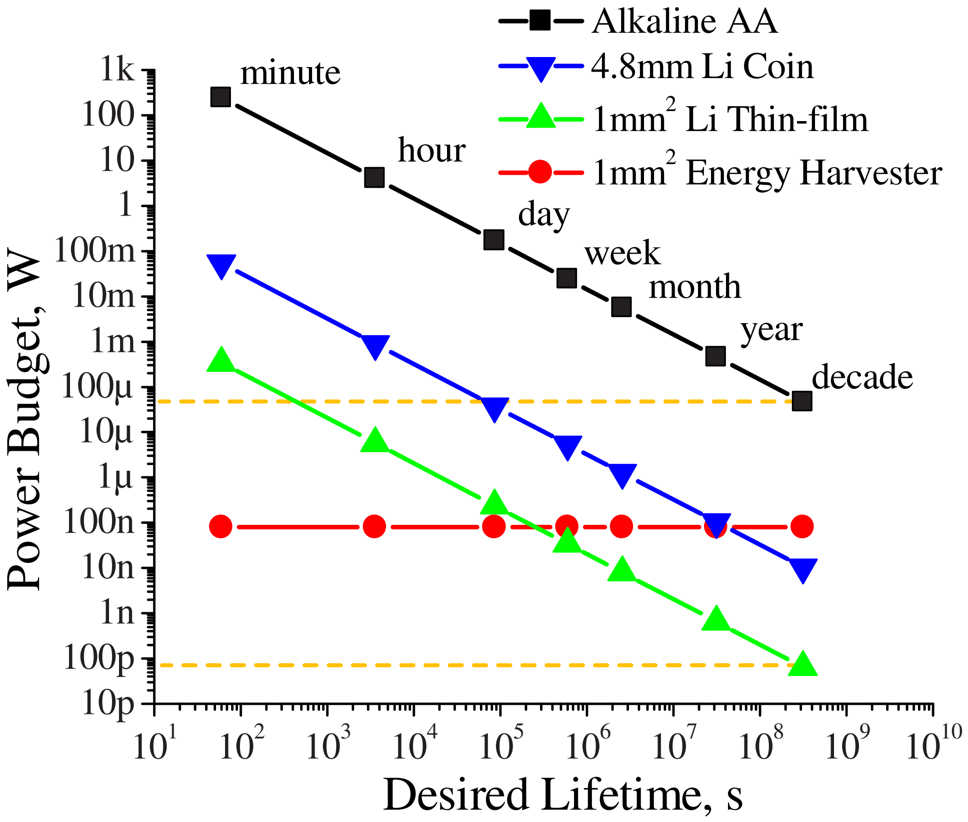


Figure 1.2: Sensor power budgets and lifetimes with common power sources.

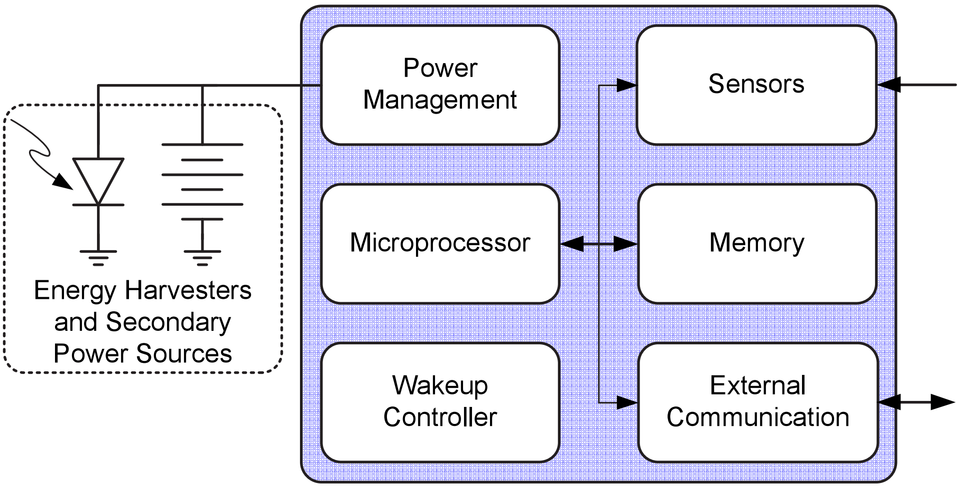


Figure 1.3: A canonical millimeter-scale sensor system.

a wired or wireless interface. Energy-autonomous sensors are powered from energy harvesters that deliver power to load circuits and recharge secondary power sources through power management units (PMUs). Previous work has recognized the need for ultra-low-power circuits that are suitable for miniature sensors. This work will be reviewed in this dissertation and includes low-power techniques, microsystem subcomponents, and complete sensor systems. During my doctoral studies, I have contributed to some of these research projects that have moved the field toward realizing millimeter-scale microsystems (Table 1.1).

In this dissertation, we will overview two millimeter-scale energy-autonomous microsystems: 1^{st} a cubic-millimeter wireless microsystem that monitors intraocular pressure (IOP) as part of a treatment for glaucoma [14]; 2^{nd} a nearly-perpetual temperature sensor that processes data with a commercial ARM® Cortex-M3™ μP [10]. These microsystems represent the first-known complete, self-powered, millimeter-scale computers. They achieve zero-net drain of battery energy by using solar energy harvesting and ultra-low-power operation. This dissertation details the power management and SRAM subcomponents of the sensor systems. Chapters 2 and 3 discuss the optimization of energy harvesters and dc-dc voltage converters for ultra-low harvester and load powers. Chapters 4 through 6 discuss robust SRAM techniques that enable nanoWatt processing for medium throughput applications such as image and audio processing in millimeter-scale microsystems [5][12][17].

Table 1.1: Research contributions to millimeter-scale sensors.

Project	**Contribution**	**Ref.**
Yield-driven near-threshold SRAM design	Simulation methodology Simulation execution	[5][6]
Reconfigurable cache	SRAM analysis	[7]
Routing for fault-tolerant NoCs	Algorithm development contributions	[8]
Hybrid Fibonacci and linear regulator DC-DC converter	Linear regulator design Testing	[9]
Millimeter-scale nearly perpetual temperature sensor system	Top-level integration PMU design Wakeup controller design Level converter design Top-level testing contributions PMU testing	[10]
Ubiquitous sensing environments	Literature review	[11]
Crosshairs SRAM - an adaptive memory for mitigating parametric failures	SRAM design BIST design Testing	[12]
Double patterning lithography and SRAM variability	SRAM design BIST design	[13]
Wireless sensing applications	Literature review	[4]
Cubic-millimeter energy-autonomous wireless intraocular pressure monitor	Top-level integration PMU design Wakeup controller design Level converter design Top-level testing contributions PMU testing	[14]
High-density portless SRAM	BIST design	[15]
Ultra-low-leakage 10T SRAM	Testing contributions	[16]
Half-differential SRAM with lower V_{MIN} and higher stability margins	SRAM design BIST design Testing	[17]

CHAPTER 2

A Cubic-Millimeter Energy-autonomous Wireless Intraocular Pressure Monitor

2.1 Continuous Intraocular Pressure Monitoring

2.1.1 Motivation

We present the first ubiquitous computer to combine sensing, processing, and wireless communication into a cubic-millimeter form factor. The microsensor monitors intraocular pressure (IOP) as part a treatment for glaucoma. It allows doctors to more accurately sense pressure levels within the eye, more quickly diagnose the severity of glaucoma, and improve their treatment regiments. Glaucoma is the leading cause of irreversible blindness, affecting 67 million people worldwide [18]. The disease damages the optic nerve due to elevated IOP and can cause complete vision loss if untreated. Currently, IOP is the only metric used to diagnose the disease and check if the patient is responding to treatments. IOP is commonly assessed using a single tonometric measurement, which provides a limited view since it fluctuates with circadian rhythms and physical activity. Continuous measurement would give doctors more information with faster response time about the severity of the disease and the efficacy of treatments. It can be achieved with an implanted monitor to improve treatment regiments, assess patient compliance to medication schedules, and prevent unnecessary vision loss.

2.1.2 Background

The ophthalmological community has recognized the benefits and challenges of continuous IOP monitoring, motivating previous works in this area [18]. Patients measure their own IOP through the eyelid without anesthesia [19]. This procedure is more convenient for frequent monitoring, but requires extensive patient intervention. A contact lens measures IOP-related changes in eye curvature [20]. This reduces patient intervention, but requires the patient to wear the contact for long periods of time including at night and is an indirect measurement of IOP. A pressure sensor is implanted as part of a replacement lens (Fig. 2.1) [21]. This provides a more direct measurement of IOP over longer periods of time than a contact can be worn, but requires lens replacement. IOP monitors can be implanted in the anterior chamber of the eye to perform long-term monitoring with minimal invasiveness and low patient intervention. A passive 4mm x 1mm monitor measures IOP with 1mmHg resolution [22]. Using an active monitor allows IOP measurements without per-measurement patient intervention for power transmission. Including a large antenna with these monitors allows them to be inductively powered and improves data transmission, but complicates implantation procedures and worsens eye trauma. A 2.3μW monitor measures IOP with 0.9mmHg accuracy and uses a 1cm PCB loop antenna [23]. Another implanted solution stores energy on a 24μF capacitor array and uses a 27mm antenna [24].

2.2 Intraocular Pressure Monitor Overview

The most suitable implantation location is the anterior chamber of the eye, which is surgically accessible and out of the field of vision (Fig. 2.1). The desired IOP monitor (IOPM) volume is limited to 1.5mm^3 (0.5$\times$1.5$\times$2mm^3) by the size of a self-healing incision, curvature of the cornea, and dilation of the pupil. The aggressive IOPM size constraint creates major challenges for achieving high-resolution capacitance measurements, wireless communication, and multiyear device lifetime. Little energy can be stored on the tiny microsystem, calling for ultra-low power operation and energy harvesting. The required millimeter antennas or inductors result in lower received power and higher transmission frequency, both increasing microsystem power.

A 1.5mm^3 microsystem is presented for long-term IOP monitoring and wireless data transmission, with self-powered energy-autonomous operation (Fig. 2.2). The 2mm x 1.5mm x 0.5mm

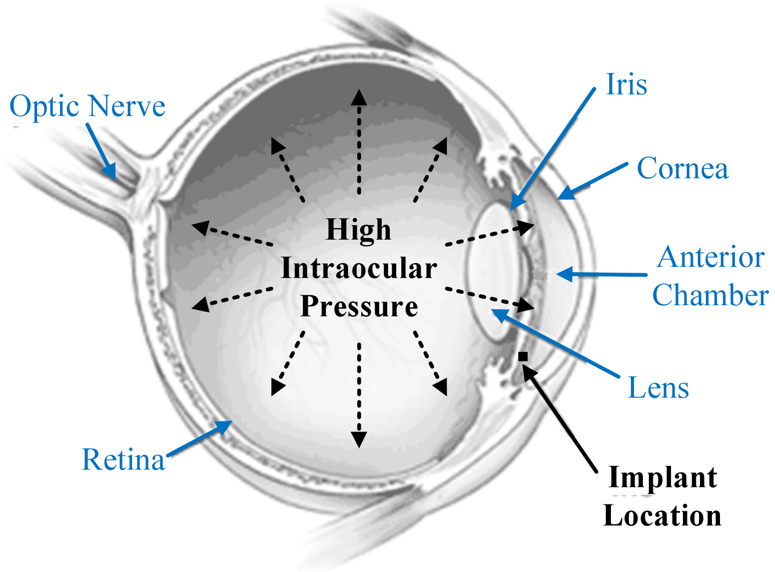


Figure 2.1: The 1.5mm^3 IOP monitor is implanted in the anterior chamber of the eye as part of a treatment for glaucoma.

monitor contains two integrated circuit (IC) chips fabricated in a 0.18μm CMOS process (Fig. 2.3). The top IC contains a solar cell and fully-integrated wireless transceiver (TRx). The bottom IC contains a sigma-delta capacitance to digital converter ($\Sigma\Delta$ CDC) sensor interface, microprocessor (μP), and static random access memory (SRAM). The solar cell and a thin-film solid-state Li battery deliver power to the microsystem through power management units (PMUs), allowing microsystem operation without patient intervention for power delivery. The ICs and battery are encapsulated in a biocompatible glass housing. A MEMS capacitive pressure sensor is anodically bonded to the outside of the housing and exposed to the ocular environment [25]. Communication and power transfer among the microsystem components uses wire bonds, with through-glass silicon vias connecting the ICs and pressure sensor.

The microsystem samples IOP every 15 minutes, which represents continuous monitoring of the slowly changing biological signal [26]. The measurement interval is set with a programmable leakage-based timer located in the bottom wakeup controller (bottom WUC) (Fig. 2.4) [9]. When the timer expires, the $\Sigma\Delta$ CDC monitors the pressure sensor's capacitance and stores a 14b digital representation of the data. After completing the CDC measurement, the μP wakes up, retrieves the recently recorded IOP data from its memory-mapped location, performs digital signal processing, and then logs its result into SRAM. Then the microsystem enters a nW standby mode until the next

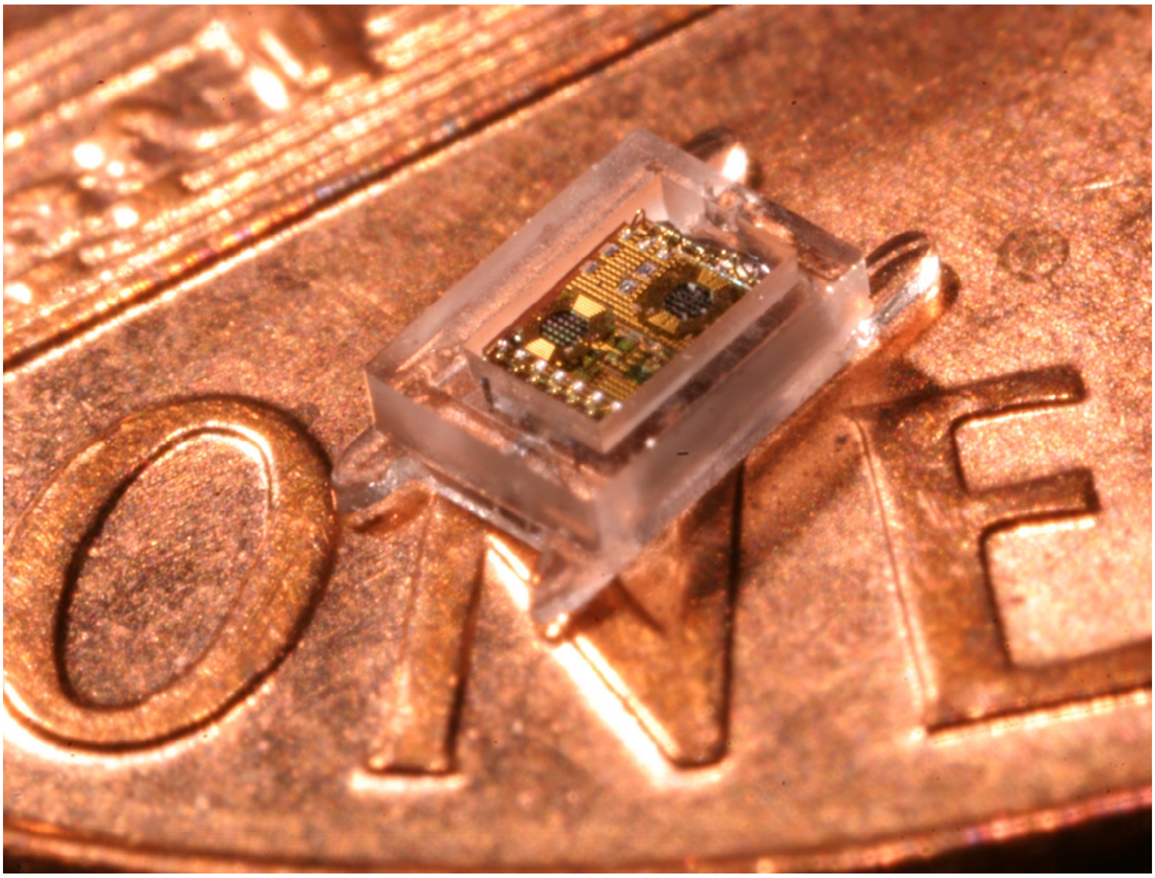


Figure 2.2: IOP monitor system photo.

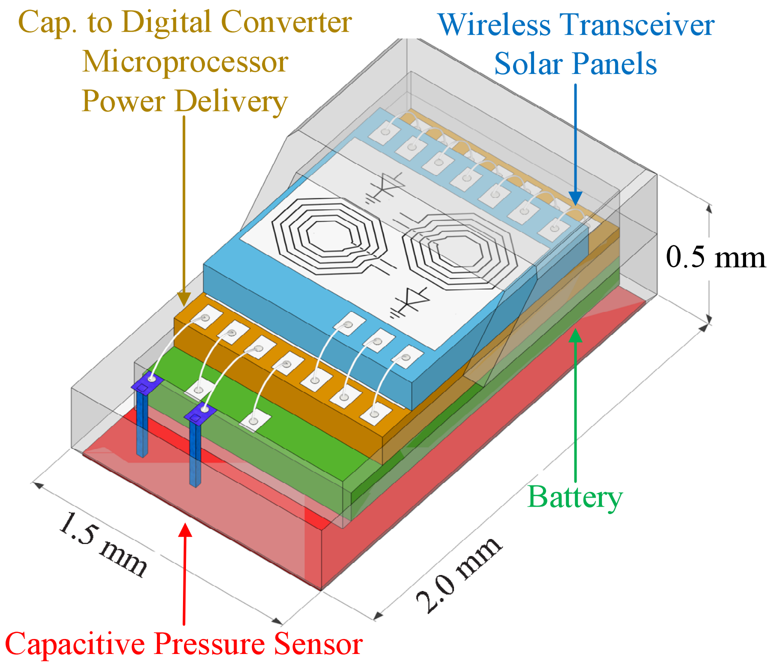


Figure 2.3: The IOP monitor includes two ICs and a battery, encapsulated in a biocompatible housing that includes a pressure sensor.

IOP measurement or user-generated event.

The implanted microsystem transmits IOP data to the user when wirelessly queried. To retrieve data from the implanted monitor, the doctor or patient places a handheld external wand over the eye. The wand transmits a wireless radio-frequency (RF) signal to the monitor. The TRx detects this signal, activates a μP program that retrieves logged data from SRAM, and coordinates the TRx as it send data to the external wand. After data transmission is complete, the microsystem again enters a nW standby mode to conserve power until the next event.

IOP monitors should achieve a lifetime of years to provide feedback to doctors while they tailor glaucoma treatments for each patient. The demonstrated microsystem operates with zero-net drain of battery energy for an indefinite lifetime using a combination of solar energy harvesting and ultra-low-power operation. The solar cell provides a 0.5V power source (V_{SOLAR}) which is used to recharge the 3.6V battery (V_{3P6}). Voltage up-conversion is performed with a fully-integrated switched capacitance voltage regulator (SCVR) located in the bottom PMU. The microsystem conserves energy by using low-power techniques. Weak inversion biasing is used in the 7.0μW CDC and short transmit pulses are employed in the 4.7nJ/bit TRx. Near-threshold processing reduces μP and SRAM active mode energy to 90nW. Power-gating, length-biasing, and device

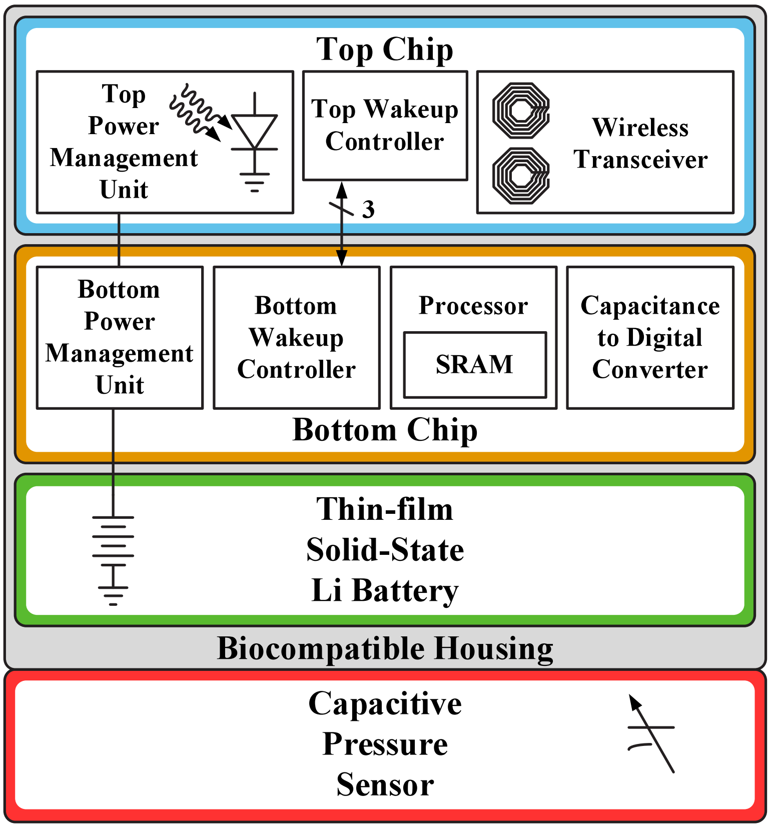

Top Chip
Top Power Management Unit
Top Wakeup Controller
Wireless Transceiver
3
Bottom Power Management Unit
Bottom Wakeup Controller
Processor
SRAM
Capacitance to Digital Converter
Bottom Chip
Thin-film Solid-State Li Battery
Biocompatible Housing
Capacitive Pressure Sensor

Figure 2.4: IOP monitor block diagram.

selection minimize μP and SRAM standby mode power to 72pW. The same SCVR used for energy harvesting delivers power to the low-voltage digital circuits (V_{0P45}) with 75% active mode and 40% light-load standby mode efficiencies.

2.3 Power Management and Energy Autonomy

2.3.1 PMU Background

Efficient energy harvesting and voltage regulation is necessary for energy-autonomous operation. Environmental energy is sporadic and must be stored on a secondary supply for continuous power. Solar cells output below 500mV [27] and secondary supplies commonly have higher voltages because of energy density or battery chemistry constraints, requiring dc-dc voltage conversion. Converter efficiency degrades at low loads because overheads do not scale down proportionally to delivered power. A low-power energy harvesting boost converter demonstrates 80% efficiency with a 100μW harvester and 45% at 1μW [28]. However, boost converters require large external inductors that do not meet the IOPM volume constraints. Switched capacitor voltage regulators (SCVRs) can be designed to perform dc-dc conversion using integrated capacitors. Harvesting SCVRs achieve 84% efficiency at 1mW [29] and 40% at 958nW [10].

Energy stored on the secondary battery must be converted to the load circuit voltage levels. A buck-converter demonstrates 80% efficiency at a 1μW load [30]. As with energy harvesting boost converters, buck-converters require external inductors which are too large for the IOP monitor, and SCVRs achieve similar functionality with integrated capacitors. A down-converting SCVR achieves approximately 76% efficiency at a 100μW load and 60% at 1μW [31]. Linear regulators exhibit high line regulation and low output noise, but suffer from poor efficiencies at high conversion ratios and cannot be reused for energy harvesting. Hybrid SCVR plus linear regulator systems increase efficiency by reducing linear regulator dropout in systems with high conversion ratios. Hybrid dc-dc converters demonstrate efficiencies of 55% at 126nW [9] and 18% at 100pW [10]. A performance summary of demonstrated previous works in energy harvesting and voltage regulation is shown in Table 2.1.

Table 2.1: Performance summary of previous works in dc-dc voltage conversion.

Reference	Method	Efficiency	Power	V_{IN}	V_{OUT}
[28]	Boost	80%	100μW	250mV	1V
[28]	Boost	45%	1μW	250mV	1V
[29]	SCVR	84%	1mW	1.15V	2.1V
[29]	SCVR	10%	1μW	1.15V	2.1V
[10]	SCVR	40%	958nW	865mV	3.6V
[30]	Buck	80%	1μW	1.2V	500mV
[31]	SCVR	76%	100μW	1.2V	500mV
[31]	SCVR	60%	1μW	1.2V	500mV
[9]	Hybrid	55%	126nW	3.6V	444mV
[10]	SCVR	41%	100nW	3.6V	400mV
[10]	SCVR	18%	100nW	3.6V	400mV

2.3.2 PMU Implementation

The PMU recharges the battery from the solar cell and delivers power to load circuits (Fig. 2.5). The 0.07 mm^2 solar cell is fabricated in an unmodified CMOS process using the deep n-well to p-well and p-well to n-active diodes. It is located underneath the transparent TRx coil openings to save area. The output of the solar cell (V_{SOLAR}) connects to the V_{0P45} node of the SCVR. When V_{SOLAR} exceeds 450mV, the 8:1 SCVR up-converts the harvested energy to above 3.6V and recharges the battery. The thin-film solid-state Li battery is supplied by Cymbet Corporation and uses a commercial chemistry, but its size is tailored specifically for this application [2]. V_{0P45} also supplies the μP and SRAM, allowing the solar cell to directly power the load circuits. When the open-circuit solar cell voltage drops below V_{0P45}, the solar cell acts as a load to the microsystem and is disconnected. The open circuit voltage is detected using a small replica solar cell. When V_{0P45} drops below 0.45V, power is down-converted from the battery by the same SCVR used for energy harvesting and delivered to the load circuits. Voltage comparisons in the PMU are performed with clocked comparators that are switched by the SCVR and top WUC clocks.

The SCVR uses an 8:1 ladder topology to perform dc-dc voltage up and down conversion (Fig. 2.6). The ladder contains dc nodes at 450mV intervals, and transfers charge up and down the ladder between adjacent nodes. Eight 35 pF MOS capacitors store charge on the dc nodes. MOS capacitors achieve higher density, and their parasitics have little effect on SCVR performance since the voltages on these capacitor terminals do not change. Seven 45 pF MIM capacitors transfer

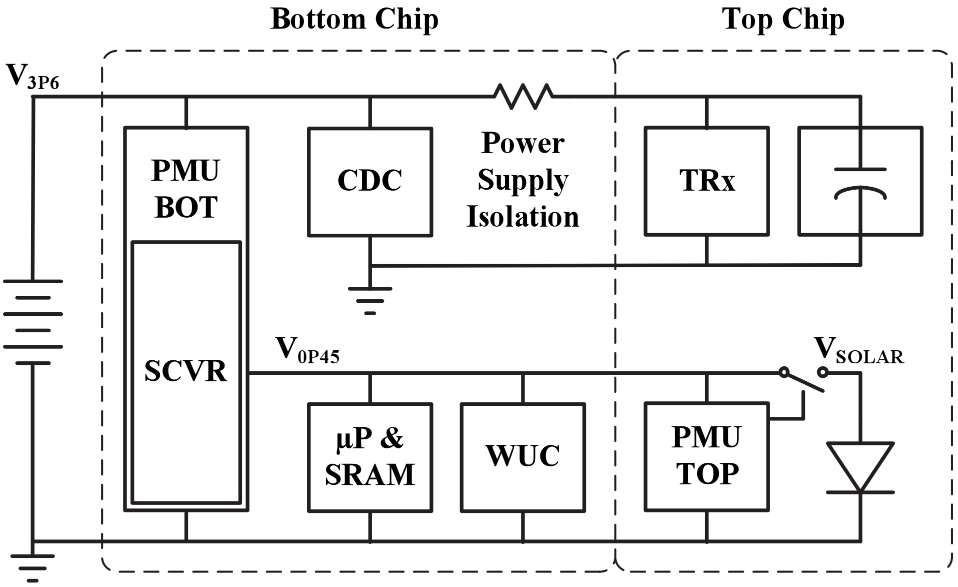


Figure 2.5: The battery is recharged from the solar cell, directly powers the CDC and TRx, and is converted down to supply the digital components.

charge between the dc nodes and are located on top of the MOS caps in the physical design. The voltages on these capacitor terminals change when the SCVR is clocked, so MIM capacitors are used to reduce parasitic bottom plate capacitance. High efficiency at nanoWatt loads requires ultra-low-power operation of SCVR circuitry. Switching losses dominate conductive losses at the IOP monitor's low power levels, so clock load and voltage swing is minimized to increase efficiency. The power switches are minimum sized high-V$_{TH}$ thick-t$_{OX}$ IO devices. NMOS switches are used except for the highest voltage portions of the ladder, where PMOS devices are used to increase overdrive. The 100kHz clock is generated at 0.9V from an internal SCVR node, level converted to 1.8V clocks with dc offsets of 0V, 0.9V, and 1.8V, and used to drive the power switches. Level conversion is performed with differential cascade voltage switch (DCVS) gates that are supplied from internal SCVR nodes.

2.3.3 PMU Measured Results

The IOP monitor is tested in a laboratory to verify functionality and energy autonomous operation (Figs. 2.7 and 2.8). This section details the measurement results for the IOP monitor's PMU. The PMU is tailored for specific voltage levels and load powers to maximize sensor life-

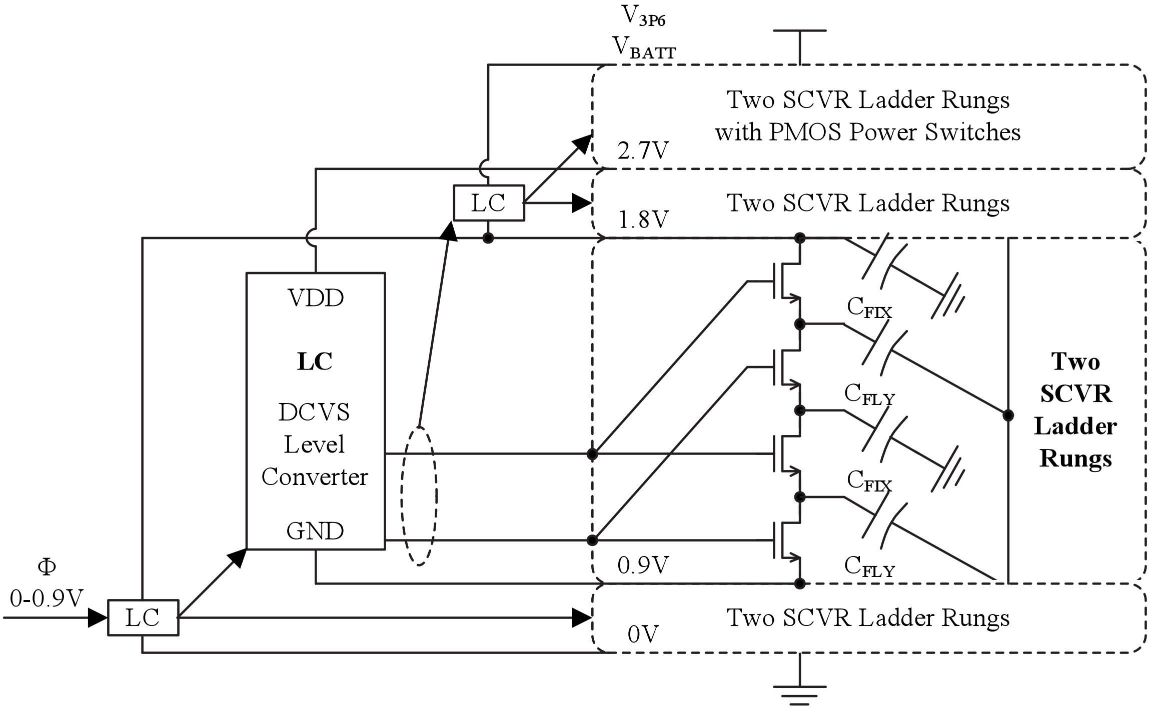


Figure 2.6: The SCVR uses an 8:1 ladder topology to perform dc-dc conversion between the solar cell, battery, and load circuits.

time. Therefore the PMU operation and optimization depends heavily on the microsystem's power sources and loads. With IOP measurements every 15 minutes, 10,000 μP cycles for DSP per measurement, and daily transmission of 14b raw IOP data, the average microsystem power is 5.3nW (Table 2.2). Standby TRx leakage and active CDC power dominate the microsystem's energy requirements. The transmitter uses large devices in the combined LO and PA to transmit sufficient power for the external wand to detect. These large devices have 3.3nW leakage in standby mode even though they are implemented as low-leakage IO devices. The CDC requires 10,000 cycles to achieve the target IOP accuracy of 0.5mmHg, setting the CDC energy per measurement. Digital components make smaller contributions to IOP monitor energy usage. In active mode, the SCVR delivers power from the 3.6V battery to the 90nW 0.45V μP, SRAM, and WUCs with 75% efficiency (Fig. 2.9). In standby mode, it delivers 72pW with 40% efficiency.

The IOP monitor achieves energy autonomy by harvesting more energy than it requires for operation. In sunlight, V_{SOLAR} reaches 500mV, and 80nW is delivered to recharge the battery (Fig. 2.10). In bright indoor lighting, V_{SOLAR} drops to near 460mV, but the battery is still recharged with 13nW, more than offsetting the 5.3nW monitor's power consumption. The IOPM requires

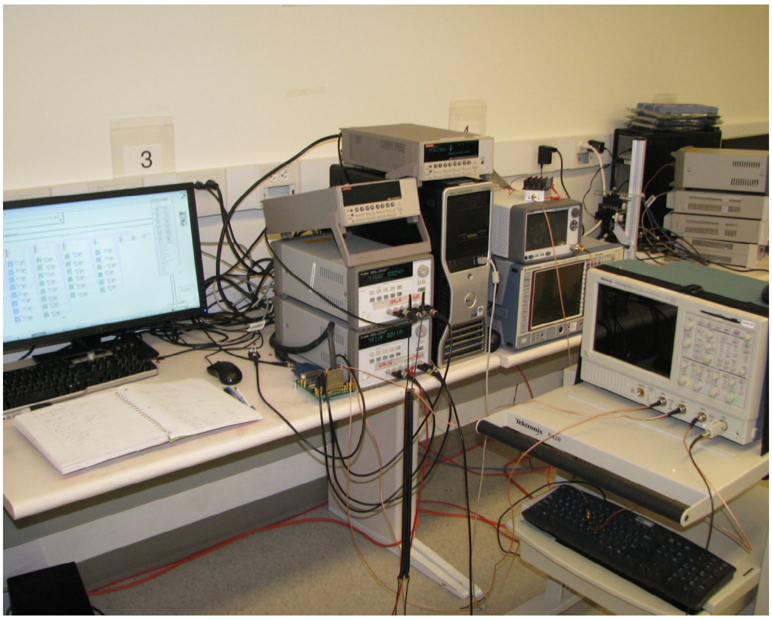


Figure 2.7: IOP monitor test setup.

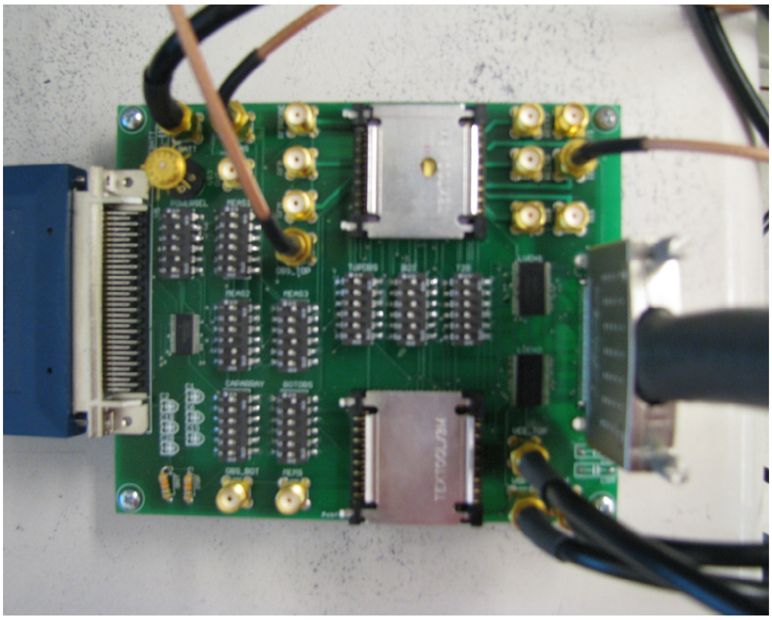


Figure 2.8: IOP monitor testing board.

Table 2.2: Power breakdown for the 5.3nW IOP monitor.

Component	**Mode**	**Power**	**Time/Day**	**Energy/Day**
CDC	Active	7.0μW	19.2s	134.8μJ
TRx	Active	47.0mW	134.4μs	6.3μJ
SCVR	Active	116.9nW	19.2s	2.2μJ
○ μP + SRAM	Active	90.0nW	19.2s	1.7μJ
CDC	Standby	172.8pW	24hr	14.9μJ
TRx	Standby	3.3nW	24hr	285.1μJ
SCVR	Standby	174.8pW	24hr	15.1μJ
○ μP + SRAM	Standby	9.8pW	24hr	846.7nJ
○ WUC	Standby	62.0pW	24hr	5.2μJ

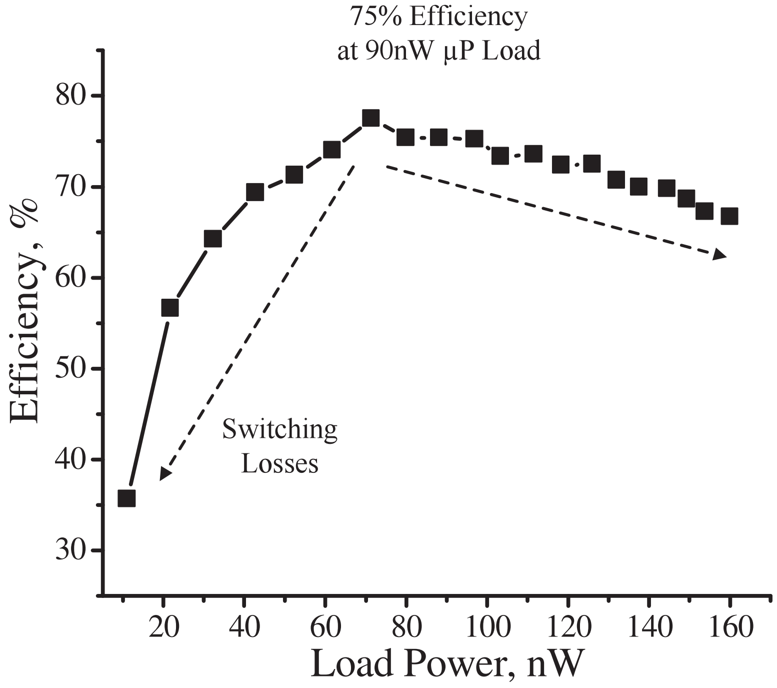


Figure 2.9: The SCVR is optimized for low switching and conduction losses to achieve 75% efficiency.

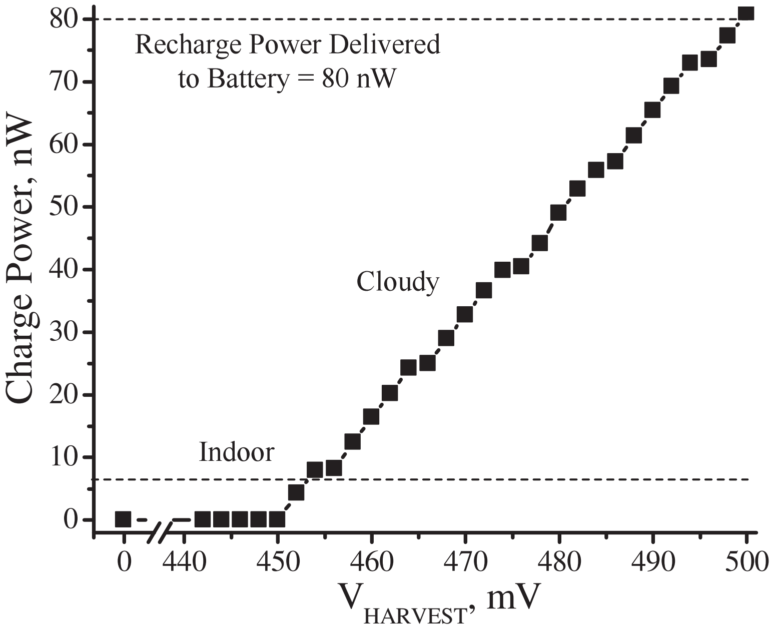


Figure 2.10: The solar cells and SCVR deliver 80nW of power to recharge the battery.

10 hours of indoor lighting or 1.5 hours of sunlight per day to achieve energy-autonomy. Energy-autonomy makes the IOP monitor lifetime decades or longer, giving doctor's enough time tailor a patient-specific glaucoma treatment.

2.4 Conclusion

We have demonstrated a complete implantable microsystem for monitoring IOP as part of a treatment for glaucoma. The 1.5mm^3 microsystem size makes it easily implantable. It measures IOP with 0.5mmHg accuracy and transmits the data through the cornea to the patient or doctor, providing rapid IOP feedback to decrease physician response time and potentially prevent unnecessary vision loss. It uses energy harvesting and ultra-low-power circuit techniques to achieve energy autonomy, extending lifetime indefinitely and giving doctors time to converge upon the best glaucoma treatment. These circuit techniques for energy autonomy enable cubic-millimeter microsystems that can be used for other medical implants, such as blood pressure and glucose sensors, as well as non-medical applications, such as supply chain and infrastructure monitoring. These pervasive sensors represent the continuation of Bell's Law, which predicts a new class of computer system's every decade and accuracy traces the evolution of computers as they have become smaller, more numerous, and more powerful. In conclusion, the development of millimeter-scale massively-deployed ubiquitous computers ensures the continuous expansion and profitability of the semiconductor industry. Energy-autonomous and nanoWatt circuit techniques will allow us to meet this next frontier in IC design.

CHAPTER 3

A Millimeter-scale Nearly Perpetual Sensor System with Stacked Battery and Solar Cells

We present the first millimeter-scale, nearly-perpetual temperature sensor. The microsystem has sensing, processing, and data storage functionality in an unprecedented form factor and power envelope. It harvests more solar energy than it consumes to perform system tasks, achieving zero-net-energy operation. The device could be used to accurately monitor working and living conditions in smart buildings, thus improving the energy-efficiency and accuracy of heating, ventilation, and air conditioning (HVAC) systems. It could also monitor pharmaceuticals during transport, to ensure that improper storage temperatures did not damage the products.

The energy-autonomous microsystem assembled into an 8.75mm^3 form factor consisting of three layers of solid-state chips (Fig. 3.1). The top layer contains two series-connected 1mm^2 solar cells which are the primary power source for the microsystem. These solar cells recharge a thin-film Lithium secondary battery on the bottom layer of the microsystem, which stores solar energy and provides instant-access power. The solar cell and battery supply power to integrated circuits in the middle layer of the assembly through an integrated PMU. These circuits include a smart temperature sensor, ARM® Cortex-M3™ microcontroller, SRAM, and wakeup controller (WUC). The microsystem harvests all of the energy required to run its low-power circuitry from the solar cells, achieving energy-autonomy. This energy is stored on a 3.6V, 12μAh, thin-film solid-state Lithium battery provided by Cymbet [2]. Energy autonomy eliminates maintenance for battery recharging or replacement, making the system more economically viable to deploy.

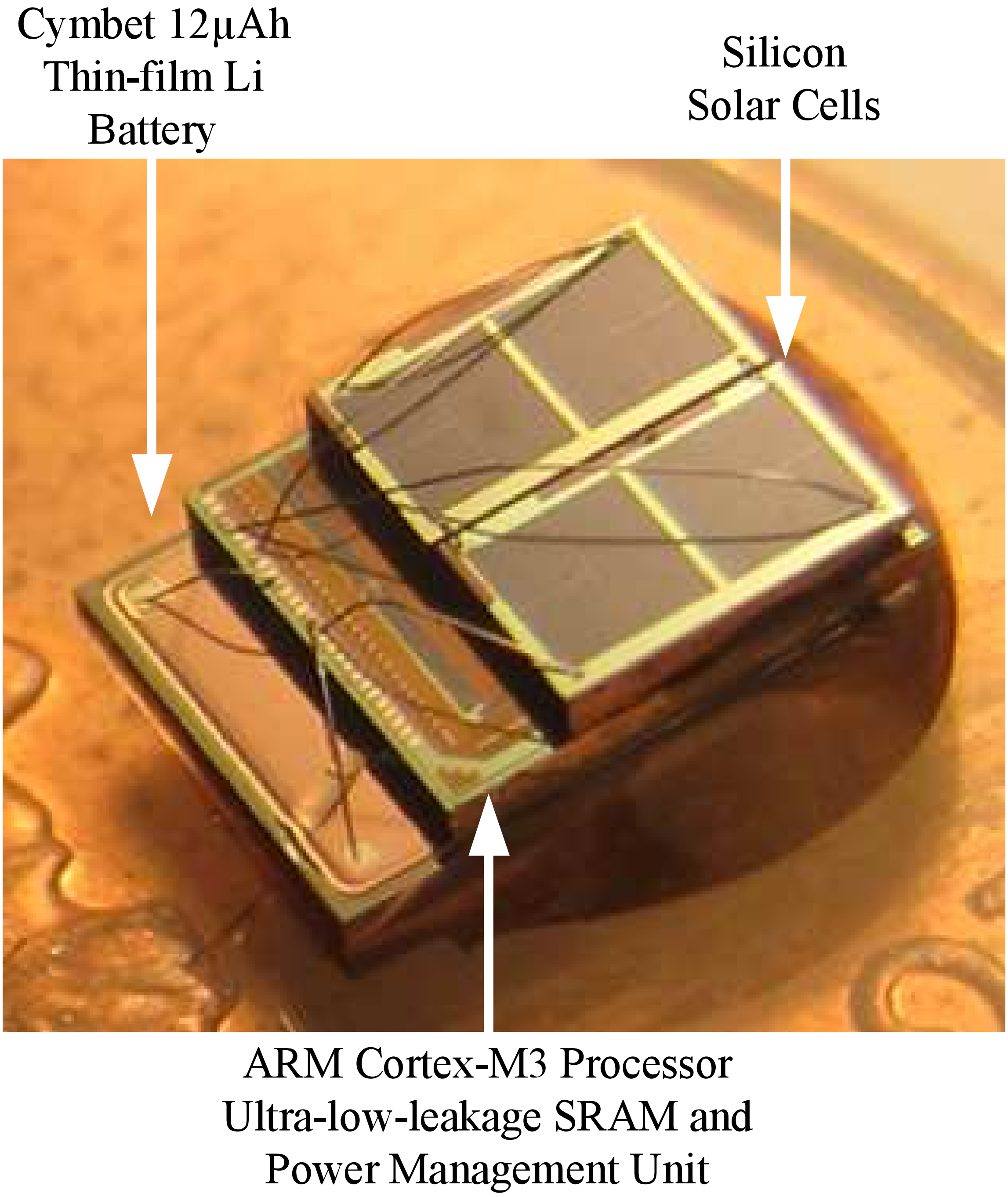


Figure 3.1: The mm-scale sensor is assembled in three layers of silicon chips for a total volume of under $9mm^3$.

The microsystem spends the majority of its lifetime in a pW standby mode and periodically wakes up to sample the environmental temperature. In standby mode, the sensor circuits and M3 microcontroller are power gated. Ultra-low-leakage 3.3fW/bit retentive SRAM (R-SRAM) holds state and data. The interval of temperature measurements is set by a programmable timer. When the timer expires, the microsystem enters active mode and the M3 reads digital temperature data. The M3 can be programmed to process and compress the data using a widely-used and trusted instruction set. Then the system logs its final result into the R-SRAM and returns to standby mode.

3.1 Background on Power Management

The power management unit in the microsystem recharges the secondary battery from the solar cells, and delivers power from the sources to the load circuits. Energy harvesters, power sources and load circuits do not operate at the same V$_{DD}$s, requiring voltage conversion. DC energy harvesters, like solar cells, are typically capable of producing output voltages at or below 500mV [27]. Since harvested energy is only sporadically available, it is stored on a secondary power supply which provides instant-access energy for the sensor. Typical secondary batteries have chemistries with output voltages ranging from 3.6V (Lithium ion) to 1.5V (Alkaline, Nickel-metal hydride, Zinc-air). The energy harvester and secondary power source may supply analog circuits requiring V$_{DD}$ above 1V and ultra-low power digital circuits near 500mV.

Voltage conversion circuits include switched capacitor networks (SCNs), low dropout linear regulators (LDOs), and Buck-boost converters. One of the primary challenges for power delivery is achieving high conversion efficiency. Buck-boost converters and SCNs ideally achieve 100% power efficiency [32]. Low-power Buck-boost converters have been demonstrated that accommodate a wide range of input and output voltages with high efficiency [28][30]. However, Buck-boost converters require a large inductor which is usually an external passive device that is larger than mm-scale. SCNs perform the same functions using large capacitors, which can be integrated on chip [31][9][29]. Low power LDOs provide improved noise, line regulation, and load regulation, which are needed by many V_{DD}-sensitive circuits [33]. However, the ideal LDO power efficiency is the inverse of the voltage conversion ratio. The efficiency equals the battery voltage (V$_{BATT}$) divided by the load V_{DD}, which is low in many battery-powered systems. In addition, LDOs perform only down conversion and are not suitable for recharging from an energy harvester.

Power management units (PMUs) for load powers ranging from mWs to tens of μWs can achieve greater than 70% efficiency, as demonstrated by a 1.2V to 0.5V, 5μW dc-dc converter [34]. However, as load currents decrease, the power of the PMU itself must scale to nW or pW levels to maintain efficiency, creating challenges for ultra-low-power SCN clock generation and distribution. A low-power 0.26mm^2 PMU demonstrates an efficiency of 55% and is suitable for load powers between 125nW and 330nW [9]. It uses a hybrid SCN plus LDO topology to capture the efficiency of the SCN and voltage stability of the LDO. However, this LDO prevents the PMU from performing up-conversion for energy harvesting. A demonstrated SCN for energy harvesting achieves 84% efficiency at 1mW and 10% at 1μW when up-converting from a 1.15V rectified source to 2.1V [29].

3.2 Power Delivery and Management

The power management unit (PMU) controls the transfer of energy between the power sources and load circuits (Fig. 3.2). This unit receives commands from the WUC, and implements the energy harvester, dc-dc voltage converter, and power gates for the microsystem. Power electronics are co-optimized for harvesting energy, recharging the battery, and down-converting power to the load circuits.

3.2.1 Solar Cells

The system harvests all of the energy it requires for operation from two series connected 1mm^2 solar cells (Fig. 3.3). Each solar cell is comprised of parallel p-well to deep-n-well and p-well to n-active p-n junctions fabricated in a 0.18μm CMOS process. No post-processing steps to improve the solar cell performance were performed. Fabrication steps built into the CMOS process were used to remove the opaque nitride and silicide layers over the solar cell and improve solar efficiency. The absence of exotic materials and specialized assembly processes reduce the cost and ease dense integration with integrated circuits. If a commercial solar cell were used, then the system could achieve energy-autonomous operation in lower lighting conditions.

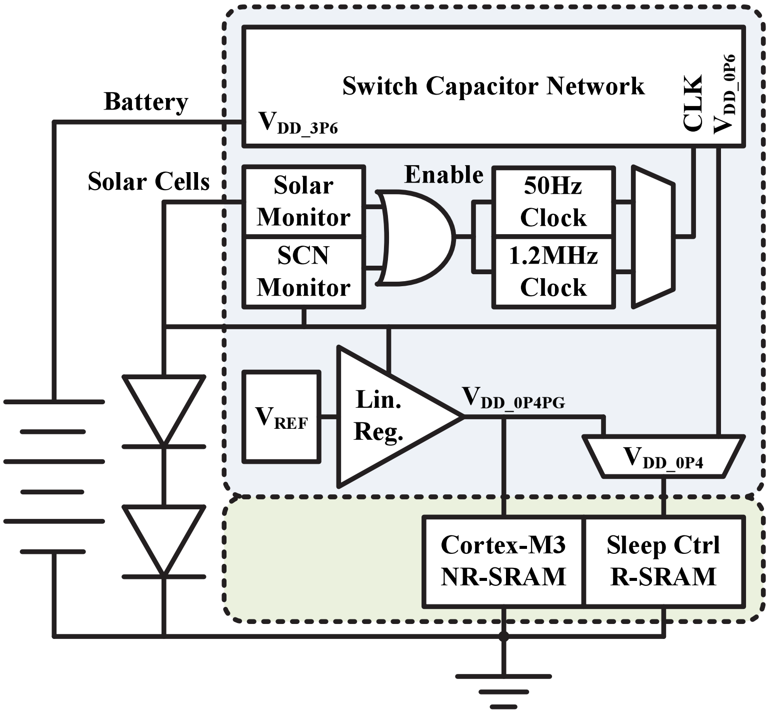


Figure 3.2: The power management unit is co-optimized to harvest solar energy, recharge the secondary battery, and efficiently deliver power to the load circuits in active and standby modes.

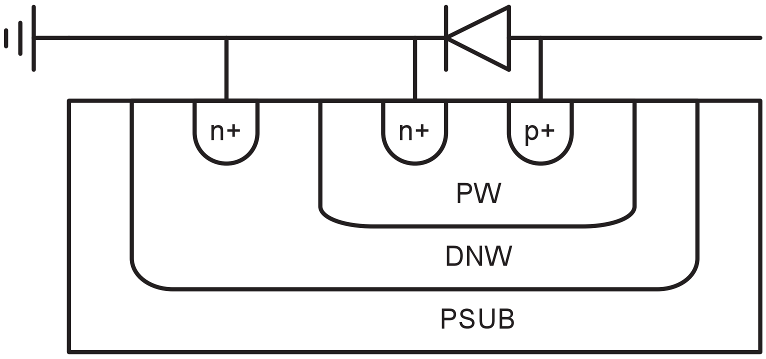


Figure 3.3: The solar cell is implemented in a standard CMOS process to reduce cost and ease dense integration with integrated circuits.

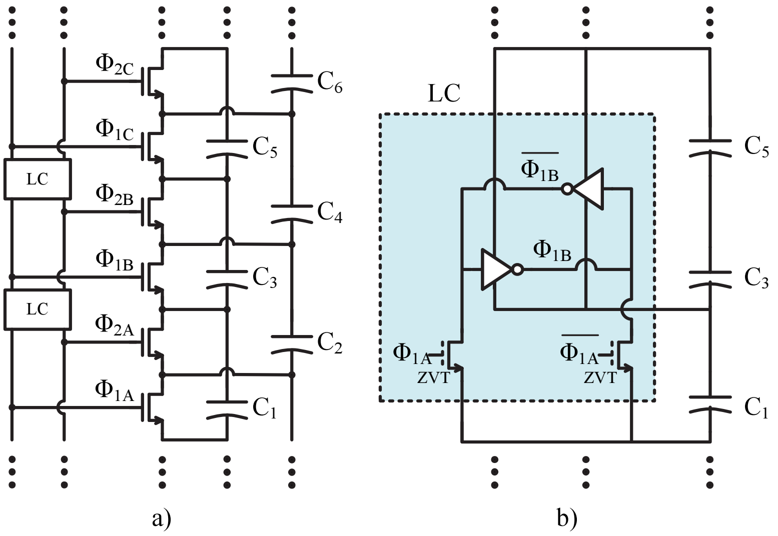


Figure 3.4: The switched capacitor network converts power at different voltages among the power sources and load circuits. Reduced swing clocks and level converters (LC) are used to reduce switching overhead.

3.2.2 Energy Harvesting

Energy is stored on a battery since the system requires a constant V$_{DD}$ but solar energy is sporadic. The PMU uses a ladder-topology switched capacitor network (SCN) to up-convert the voltage from the two series PV cells (<900mV) by 6x to recharge the battery (3.6V) (Fig. 3.4). Six series metal-insulator-metal (MIM) capacitors provide charge storage for internal SCN nodes, with voltages from 0V to 3.6V in steps of 0.6V. Another 5 MIMs are switched to transfer charge between successive nodes. For our target load, SCN switching losses are more dominant than conductive losses. Therefore, the SCN reduces clock overhead by using 1.2V clocks to drive its power switches, instead of full-swing 3.6V clocks. The PMU clock generator is generated from the 1.2V internal node of the SCN and level converted up the ladder. Level converters have a cross-coupled inverter pair supplied from a higher internal SCN voltage, and native zero-V$_{TH}$ write pass transistors that are driven by lower voltage clock signals.

When there is no light, the solar cells are disconnected from the PMU, since they act as large forward-biased pn-junction and load to the SCN. The lighting condition is calculated by monitoring the open-circuit voltage (V$_{OC}$) on replica solar cells (RSC) included in the system. The voltage on the harvesting solar cells (HSC) cannot be monitored since they may be shorted to the SCN. The total solar cell area seen in Fig. 3.1 is 4mm^2, with 2mm^2 HSCs and 2mm^2 RSCs. When V$_{OC}$

exceeds V$_{DD\_0P6}$, the HSCs are connected to the SCN and input energy into the microsystem. The HSC voltage is compared to V$_{DD\_0P6}$ with a clocked comparator using the SCN clock.

3.2.3 Standby Mode

Reducing total energy consumption requires reducing the power of the sensors, microprocessor (μP), and SRAM and also efficiently delivering power to these load circuits from the power sources. Voltage down-conversion from the battery is performed with the same SCN used for energy harvesting. The SCN automatically converts up from the solar cells or down from the battery based on the power source and load V_{DD} levels.

For an SCN with these ultra-low-power loads, the power overhead must scale to pW and is dominated by clock generation and distribution. Switching power is reduced by minimizing the clock frequency and load. The SCN frequency is dynamically scaled from a maximum of 50Hz. The 50Hz SCN accommodates the worst case load current for standby mode. However, loads vary with temperature and system activity. When load current decreases, slower SCN operation will adequately power the system with lower overhead. To achieve this, the SCN is only clocked when V$_{DD\_0P6}$ drops below an acceptable voltage of 550mV. V$_{DD\_0P6}$ is compared with a reference voltage using a clocked comparator, whose output is used to gate the SCN clock. This on-demand power delivery yields a 20% improvement in efficiency.

Generating a pW ultra-low-frequency clock is one of the primary WUC design challenges. A clock can be generated using a current-started ring oscillator (CSRO) with an analog bias voltage. However for very low frequencies, these structures have reduced output swing and high frequency sensitivity to the bias voltage value. Alternatively, we use a leakage-based ring oscillator (LRO) to alleviate these issues.

Fifteen 0.4V leakage-based delay elements are connected into a ring to create the 50Hz LRO. Each delay element switches between two stable states using leakage power (Fig. 3.5). Starting in one stable state, the differential output nodes (Z, Zbar) are held to the power rails. After the differential inputs (A, Abar) transition, the outputs float. Differing drain to source voltages across the leaking devices will cause the output nodes to drift toward the metastable state of the cross-coupled pair. This changes the gate voltages on the leaking transistors and initiates a positive

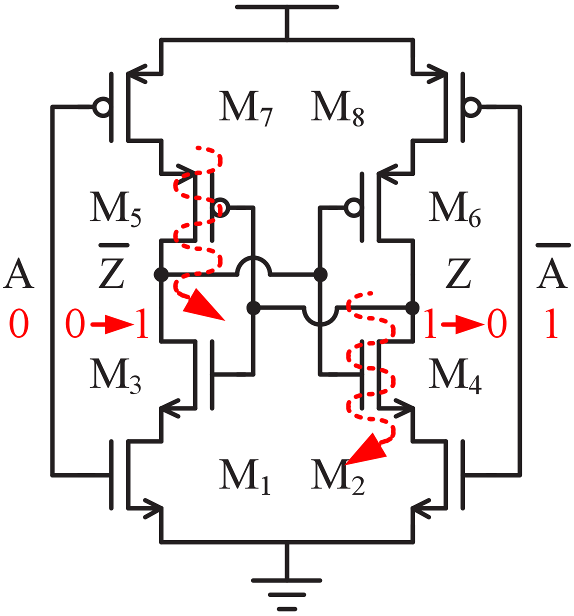


Figure 3.5: Slow clocks for the wakeup controller and standby mode switched capacitor network are efficiently generated using a leakage-based delay element.

feedback mechanism to fully switch the output nodes. This leakage based operation with positive feedback creates outputs with long delay but with risetimes and falltimes an order of magnitude better, when compared to a CSRO with the same energy and frequency.

The power switches and level converters represent SCN clock loads and are minimized to reduce power. Minimum-sized level converters have adequate performance for the 1.2MHz active-mode operation. Typically, SCN switches are made very large to reduce their resistance and minimize conductive losses. However since our load currents are small, these conductive losses are negligibly small and the switches are minimum sized.

3.2.4 Active Mode

In active mode, the μP load power increases from pWs to μWs. During this transition, the power gated V_{DD} rails are recharged causing a power spike. Special care is taken to ensure this does not overly tax the PMU, causing destructive V_{DD} drops that could lead to data loss. To switch to active mode, the SCN first switches to a separate 1.2MHz clock, maximizing the power the SCN can deliver. Secondly, the power-gated V_{DD} rails are slowly charged through a resistive network (Active && !Strong in Fig. 3.6). This reduces peak power to levels the SCN can accommodate.

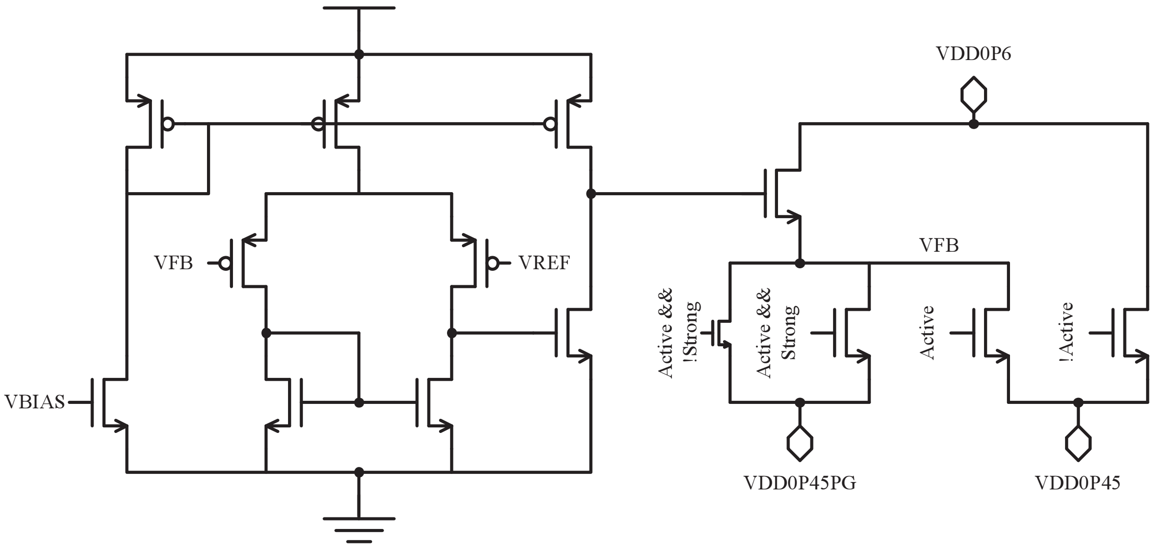


Figure 3.6: The linear regulator turns the switched capacitor output into a stable 0.4V supply across battery voltage and temperature.

Finally, after the power rails are fully charged, they are connected to the SCN though low-resistance devices to minimize power supply noise and μP operation begins.

In active mode, an LDO is added to remove voltage transients on the system supply rails due to SCN clock injection and load current transients (Fig. 3.6). To reduce LDO losses, the dropout is kept as small as possible and the LDO bias current is minimized. The LDO converts V_{DD_0P6} which is supplied by the solar cells and SCN, to 0.4V. The output voltage is set by a 2.2-pW voltage reference (Fig. 3.7) [35]. The LDO is biased with a 30nA device in weak inversion to minimize energy while still preventing unacceptable V_{DD} drops due to load transients. In standby mode, this LDO is bypassed, with its bias current eliminated. The SCN alone can provide a stable V_{DD} in standby mode since there is little switching activity and thus no sudden changes in load current.

3.3 Measured Results

3.3.1 Solar Cell

The solar cell IV curve is measured using a solar simulator to match the sun's spectrum with an irradiance of 1 sun in the air mass 1.5 standard (AM 1.5). The short circuit current is 31μA and its open circuit voltage is 540mV for one solar cell at 1 sun (Fig. 3.8). The maximum solar efficiency

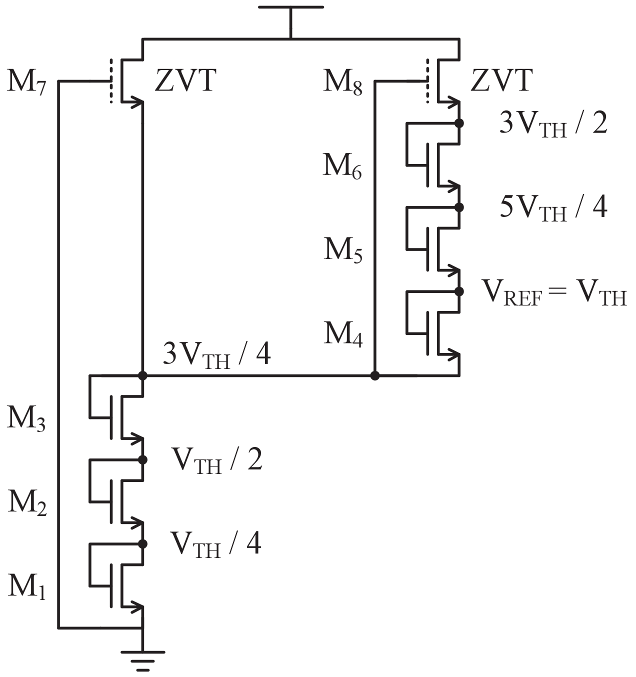


Figure 3.7: The output of the linear regulator is generated using a picoWatt voltage reference.

is 5.5%. AM 1.5 is commonly used to characterize solar cells and one sun was originally chosen to represent ideal conditions in the United States, however actual outdoor lighting intensity and spectrum can vary significantly [36]. Solar current scales almost linearly with light irradiance and our measured indoor lighting conditions are near 0.01 suns. The performance of the solar cell is poorer than commercial devices because of our selection of cheap and widely available materials. However even with a non-optimal solar cell, the microsystem achieves zero-energy operation.

3.3.2 Power Delivery

The nearly-perpetual temperature sensor system is tested in a laboratory to verify functionality and energy autonomous operation (Figs. 3.9 and 3.10). This section details the measurement results for the temperature sensor system's PMU. The hybrid voltage converter captures the energy efficiency of the SCN and voltage stability of the LDO. Using an LDO alone would result in an ideal efficiency of 11%, while the hybrid topology achieves 40% efficiency. The SCN output itself has a voltage ripple of 50mV because of SCN switching. The LDO reduces this ripple to under 1mV for DC load currents. The SCN output voltage decreases with decreased V_{BATT} and increased load currents. However, the V_{DD} is held at a constant 0.4V using the LDO. A load regulation of

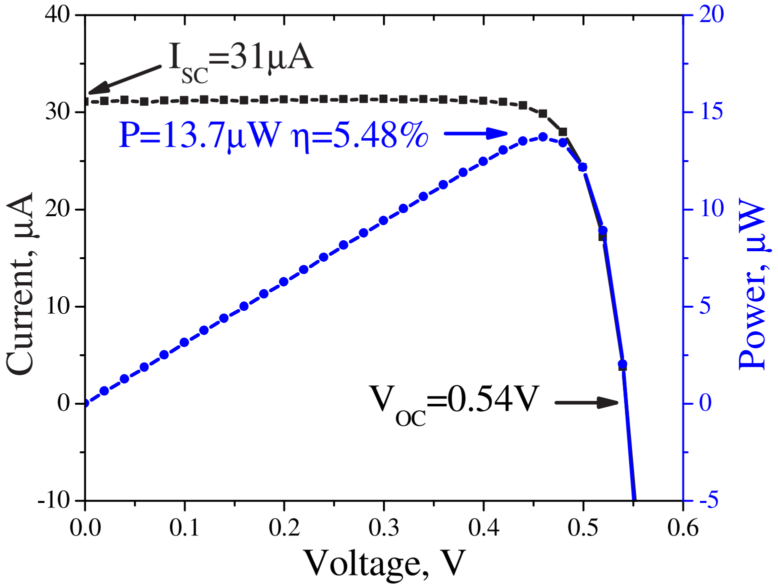


Figure 3.8: Solar cell VI curve at a light intensity of 1 sun AM 1.5.

2.4% is achieved between 0μW and 6.25 μW loads (Fig. 3.11). Also, a line regulation of 0.72% is achieved when V$_{BATT}$ degrades from 3.6V to 3.3V (Fig. 3.12).

Active mode PMU down-conversion efficiencies are measured for load currents ranging from 100nW to 2.5μW (Fig. 3.13). The SCN efficiency nears 40% at load currents above 1.5μW, with an ideal efficiency of 66% because of LDO dropout losses. Using a fixed SCN clock frequency, the efficiency degrades as load current decreases because of the fixed energy overheads in the SCN. Switching overhead is reduced using on-demand power delivery that skips SCN clock pulses when the output voltage conditions are already satisfied (3.14). This increases the low-load efficiency of the PMU, with a 4.75x efficiency improvement at a 100nW load (Fig. 3.13). With energy-efficient voltage conversion plus the ultra-low power consumption, the microsystem power is 7.7μW in active mode and 550pW in standby mode. This translates to a device lifetime exceeding 5 years with hourly temperature measurements using 10,000 μP cycles when the PMU cannot harvest light energy.

3.3.3 Nearly-Perpetual Operation

During sensor measurements and processing, the microsystem draws power from the battery because the tiny solar cells cannot deliver the required μW load powers. However, between sensor

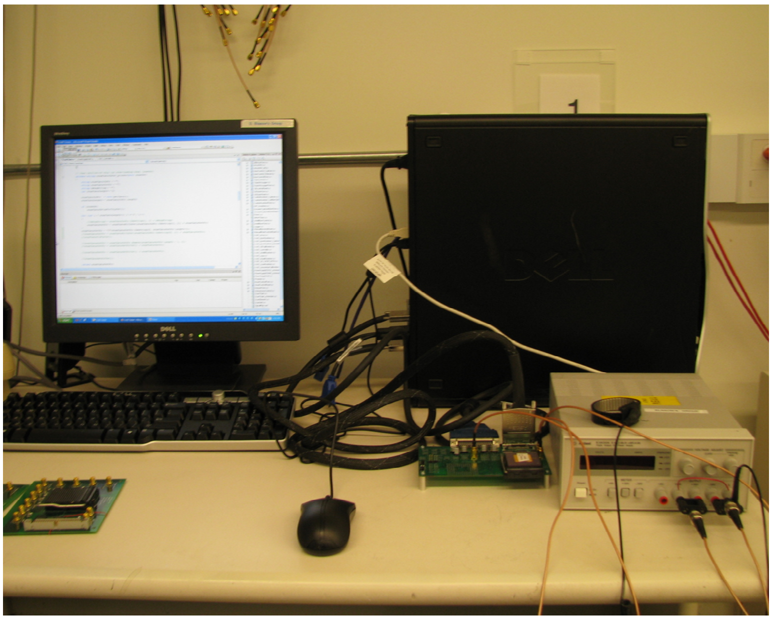


Figure 3.9: Nearly-perpetual temperature sensor system test setup.

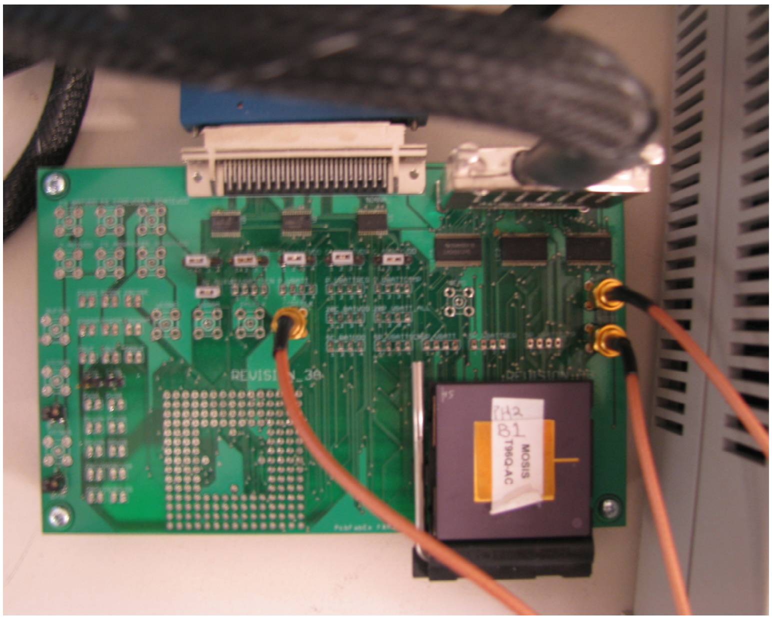


Figure 3.10: Nearly-perpetual temperature sensor system testing board.

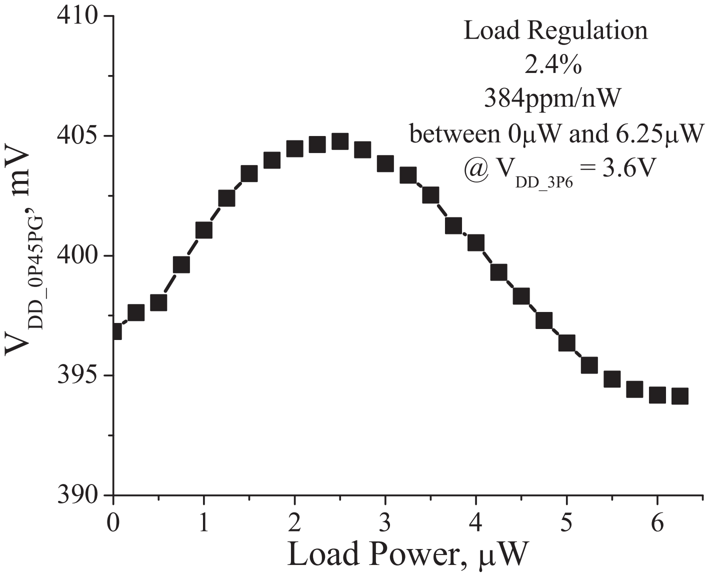


Figure 3.11: Hybrid switched capacitor network and low dropout regulator load regulation.

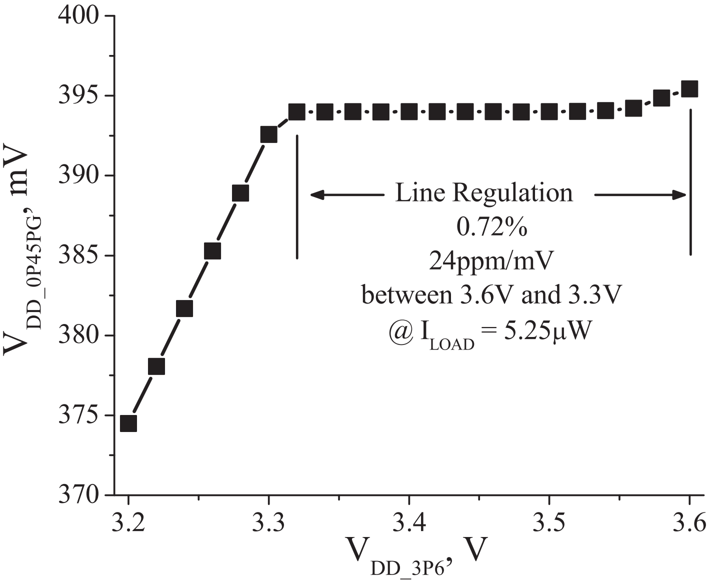


Figure 3.12: Hybrid switched capacitor network and low dropout regulator line regulation.

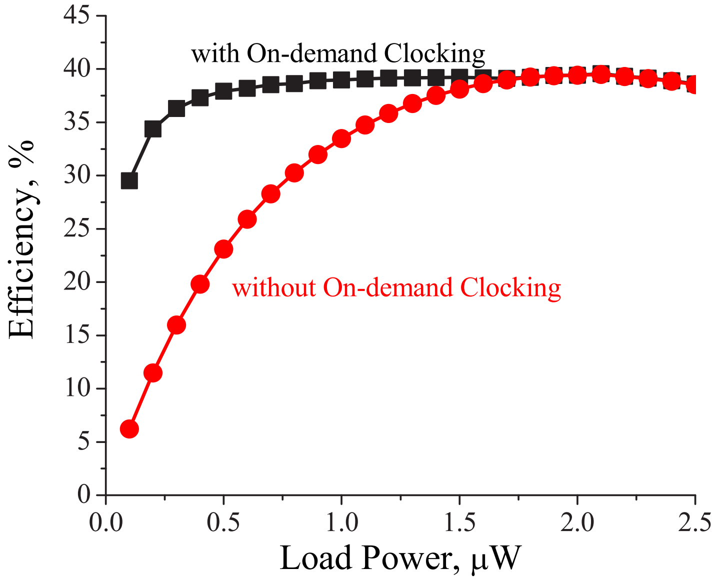


Figure 3.13: Active mode SCN down-conversion efficiency nears 40% out of an ideal 66%. Low-load efficiency is improved using on-demand power delivery.

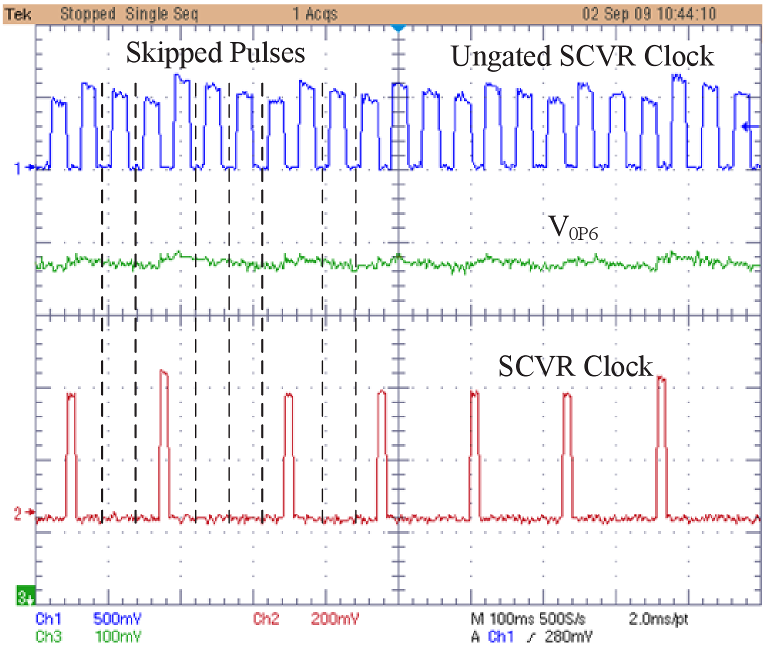


Figure 3.14: Measured waveforms of on-demand clock of the switched capacitor network, which is not switched when the output voltage is sufficiency high.

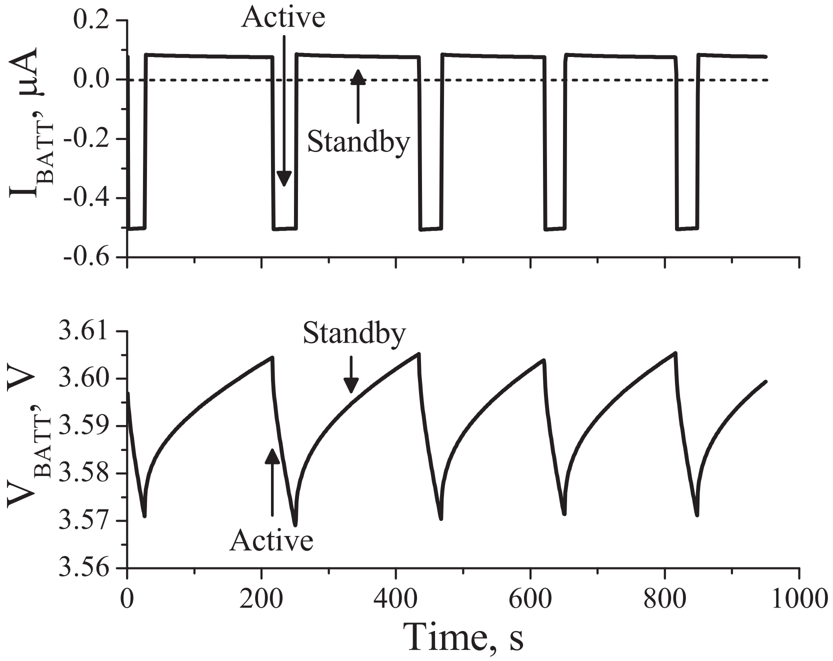


Figure 3.15: Measured battery VI curves for microsystem charging and discharging cycles.

measurements, the load power drops to pWs, and the PMU recharges the battery in both indoor and outdoor lighting conditions. Measured battery voltage and current during charging and discharging cycles are shown in Fig. 3.15. The tiny, low-efficiency solar cells harvest more solar power in indoor lighting than the microsystem consumes on average because of the low-power techniques employed. Therefore when light is available, the system lifetime does not depend on battery capacity or energy consumption but instead on device wearout.

Nearly-perpetual operation is achieved when the standby period is long enough to recuperate energy used in active mode. The necessary standby duration for energy-autonomous operation depends on the battery recharge rate, which in turn depends on light intensity. Fig. 3.16 shows the number of sensor measurements that can be taken each day in different lighting conditions with zero-net drain of the battery. This plot assumes that each sensor measurement requires 10,000 μP clock cycles to perform. Even in indoor lighting conditions as low as 0.005 suns AM 1.5, nearly-perpetual operation is achieved. On a sunny day 15,000 sensor measurements can be taken with no net energy discharge from the battery.

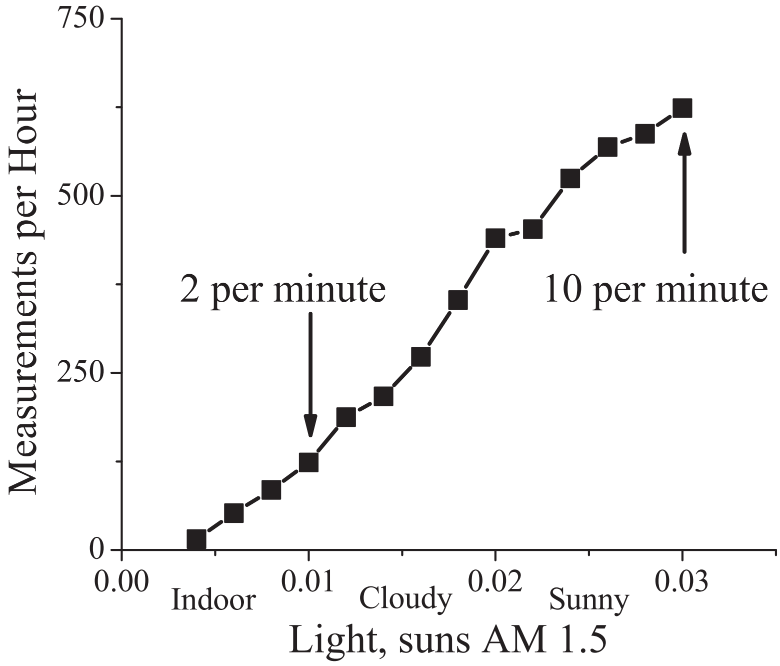


Figure 3.16: The duty cycle of the microsystem while achieving zero-net drain of the battery across lighting conditions.

3.4 Conclusion

An 8.75mm^3 temperature sensor system achieves zero-net-energy operation with energy harvesting and ultra-low-power techniques. A 28pJ/instruction, 0.4V, 72kHz ARM® Cortex-M3™ microcontroller processes temperature data. A custom 3.3fW leakage per bit SRAM stores the μP result. Two 1mm^2 solar cells and a thin-film Li battery power the microsystem through an integrated power management unit. The complete microsystem consumes 7.7μW when active. It then enters a 550pW data-retentive standby mode between sensor measurements. The microsystem can process hourly temperature data for 5 years using only energy stored on the battery. This lifetime is extended indefinitely using energy harvesting to recharge the battery, enabling nearly-perpetual operation. The microsystem's small size makes it suitable for temperature sensing in heating, ventilation, and air conditioning systems, supply-chain monitors, and many other applications.

CHAPTER 4

Yield-driven Near-threshold SRAM Design

4.1 Introduction

Intelligent millimeter-scale microsystems require memory to perform on-sensor processing. Greater memory capacities allow microsystems to store more environmental data, which is especially useful for high-throughput sensing modalities like audio and image sensing. More memory also enables more advanced processing algorithms to extract useful information from the raw sensor data. These programs include word recognition and line detection algorithms. For nanoscale ubiquitous computers, memory capacities are limited by area, and also by the energy budget. Therefore dense, low-power SRAM is necessary for intelligent cubic-millimeter microsensors.

A simple and effective way to reduce energy is to scale down V$_{DD}$. This delivers a quadratic savings in dynamic energy consumption and a linear reduction in leakage power [37] [38] [39]. As shown in Figure 4.1, as V_{DD} is scaled down into the near-V$_{TH}$ region, between 400mV and 700mV, the energy per operation is significantly reduced and delay degrades gracefully [37] [38]. As V_{DD} is scaled further, delay increases dramatically and total energy per cycle increases because leakage energy dominates. Leakage energy per computation increases as V_{DD} is scaled down, even though leakage power lowers, since it is proportional to delay, which increases exponentially in the sub-V$_{TH}$ region. There exists a V$_{DD}$ where the total energy per operation is minimized (V_{MIN}). V_{MIN} depends heavily on the ratio of dynamic to leakage energy for the circuit. Compared to combinational logic, which commonly has sub-V$_{TH}$ V_{MIN}, caches have more idle circuitry and a lower activity rate. This increases the ratio of leakage to the dynamic energy and subsequently

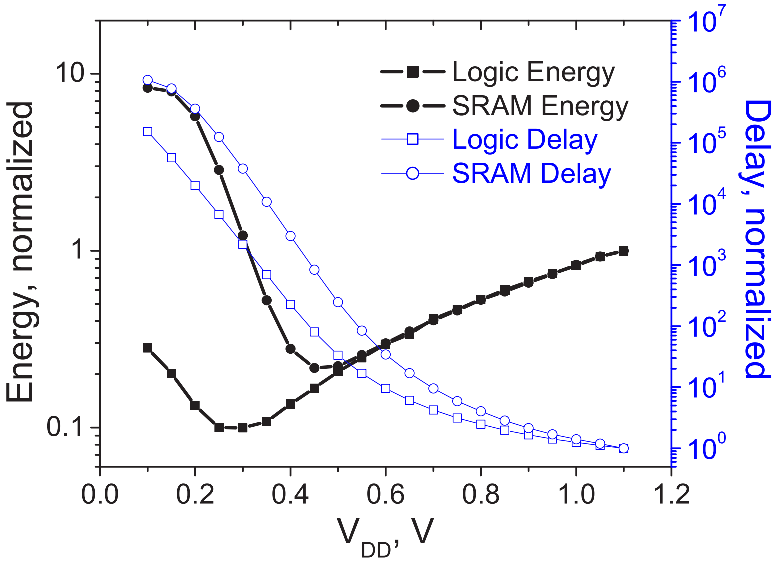


Figure 4.1: V$_{DD}$ scaling quadratically reduces dynamic energy and linearly reduces leakage power. Since delay increases exponentially, leakage energy per instruction increases and dominates total energy at low V_{DD}. These competing trends result in a V_{DD} that minimizes total energy per instruction, denoted as V_{MIN}.

increases V_{MIN} into the near-V$_{TH}$ region for common cache configurations. In this paper we target SRAM designs to robustly operate near V_{MIN} in the near-V$_{TH}$ region.

As V_{DD} is scaled down, the ON/OFF current ratio for devices is reduced and noise margins diminish. Typically, CMOS circuitry maintains adequate robustness in the presence of these effects. However, Static Random Access Memory (SRAM) becomes more prone to functional failures at low V_{DD}, as evidenced by the reduction in the static noise margin (SNM) shown in Figure 4.2.a [40]. In addition at low V_{DD}, bitcells are more susceptible to V_{TH} variation caused by random dopant fluctuation (RDF). RDF shifts the V_{TH} of each transistor independently, causing mismatch within bitcells and greatly reducing SNM, as shown in Figure 4.2.b. When SRAM V_{DD} is scaled down these V_{TH} shifts have a greater impact on device currents. At near-V$_{TH}$ V$_{DD}$s, RDF is the dominant form of process variation and the foremost reason for poor robustness in sub-V$_{TH}$ and near-V$_{TH}$ SRAM. Making SRAM transistors larger can increase SRAM robustness since non-uniformities in channel doping average out, resulting in more uniform device V_{TH}s [41]. The cost of increased device sizing is larger SRAM area and higher energy.

One proposed solution for near-V$_{TH}$ SRAM is the 8T bitcell [42]. The 8T bitcell connects two

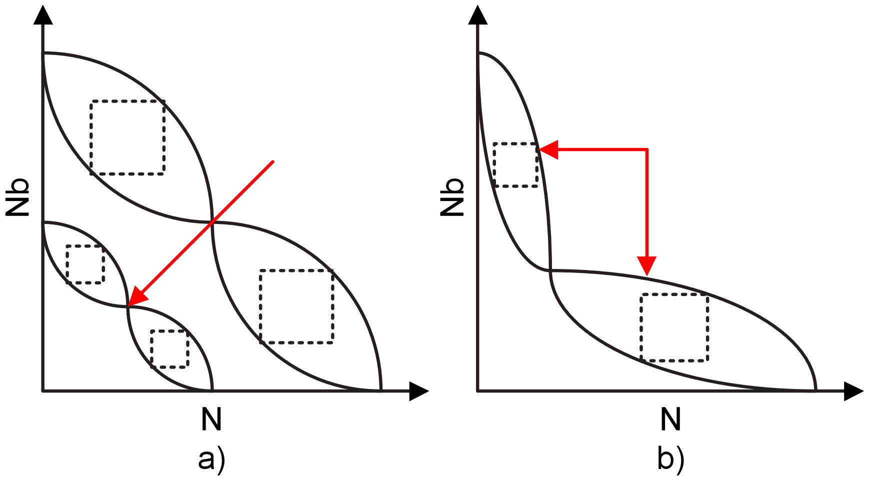


Figure 4.2: a) Static noise margin (SNM) scales with V_{DD}. b) Random dopant fluctuation (RDF) induced V_{TH} variation causes mismatch in SRAM bitcells, reducing robustness and SNM . At near-V$_{TH}$, SRAM robustness is more sensitive to V_{TH} variation because drain current is more sensitive to gate overdrive.

additional stacked NFETs to the differential 6T structure to isolate the read and write accesses. A separate read wordline and read bitline are employed to perform a single-ended read on the bitcell with no risk of upsetting the bitcell's value. The separate read structure allows the other six devices to be sized and doped appropriately to ensure write stability. A typical 8T bitcell is over 33% larger than a differential 6T bitcell but may have higher array efficiency [42]. Further solutions for high robustness SRAM use read and write assist circuits [43] [44] [45] [46] [47]. These circuits modulate the WL, BL or V$_{DD}$s to prevent functional failure. These techniques have the advantage of keeping bitcell density high, but may require extra overhead such as additional peripheral devices or voltage sources.

Numerous ultra-low energy SRAMs reduce energy by scaling V_{DD} to sub-V$_{TH}$ levels [47] [48] [49] [50] [51]. A single-ended 6T SRAM has been demonstrated that is functional below 200mV with a 40% area penalty [47]. A multiplexer tree can be used to read data values and improve read stability [48]. A 10T cell bitcell was designed with assist circuitry to improve bitline sensing [49]. Incorporating a Schmitt trigger into the cross coupled inverters can prevent read failures and improve hold margins [50]. Also, multiple-V_{TH} designs exist for improving robustness and reducing leakage [51]. Many of these ultra-low energy SRAMs exhibit insufficient robustness for commercial designs, where SRAM sizes reach MBs, limiting them to small arrays and sensor applications. Also, increased delay, and thus leakage, for these architectures may cause sub-V$_{TH}$ operation to be suboptimal for minimum energy operation.

In our study we take a new look at the existing differential 6T and 8T bitcell architectures by thoroughly comparing the two designs in robustness, area, delay, and energy in the super-V$_{TH}$ and near-V$_{TH}$ voltage regimes in order to explore energy savings through voltage scaling [42]. In our study we constrain all bitcells at all V_{DD}s to have equal robustness to a commercial differential 6T at a 1V supply. As V_{DD} is scaled down, either the bitcell doping and sizing will be adjusted or assist circuits will be employed to meet these constraints.

To calculate robustness, we model RDF-induced random process variation. The effects of process variation may be measured through either SNM measurement, corner case analysis, Monte Carlo simulation or analytical modeling [40] [52]. However, SNM analysis does not consider the dynamic nature of noise injection. Corner case analysis is pessimistic, resulting in over-optimized bitcells and unnecessary area and power. Monte Carlo simulation is extremely computationally intensive for SRAM because the acceptable failure rate is low. Alternatively, we calculate SRAM robustness using importance sampling. We sample heavily in the failure region of interest, reducing the number of samples needed to characterize the failure modes [53]. The resulting samples are weighted using device V_{TH} probabilities to calculate the SRAM yield [54]. Importance sampling allows us to accurately and efficiently calculate bitcell yield.

We find that halving V$_{DD}$s from 1V to 500mV for differential 6T bitcells halves dynamic energy with either a 40% area overhead or a 200x delay penalty for maintaining robustness. Halving V$_{DD}$s for 8T bitcells also halves dynamic energy with no area overhead and preserved cache latency. The 8T bitcell can be further scaled to 300mV to cut dynamic energy by 83% with a negligible area overhead. Using this information we find the V_{MIN} and energy at V_{MIN} (E_{MIN}). V_{MIN} can be as low as 300mV for 8T L1 caches with high access rates and as high as 950mV for L2 caches with low access rates.

In this paper we contribute a framework for selecting an appropriate SRAM architecture given a set a design constraints, including near-V$_{TH}$ robustness. We show that V_{MIN} for SRAM is significantly higher than V$_{DD}$s targeted in previous designs, and hence guide the focus of new SRAM research for energy efficiency. The rest of this paper is organized as follows: Section 4.2 discusses the topology and operation of the candidate architectures; Section 4.3 examines the simulation

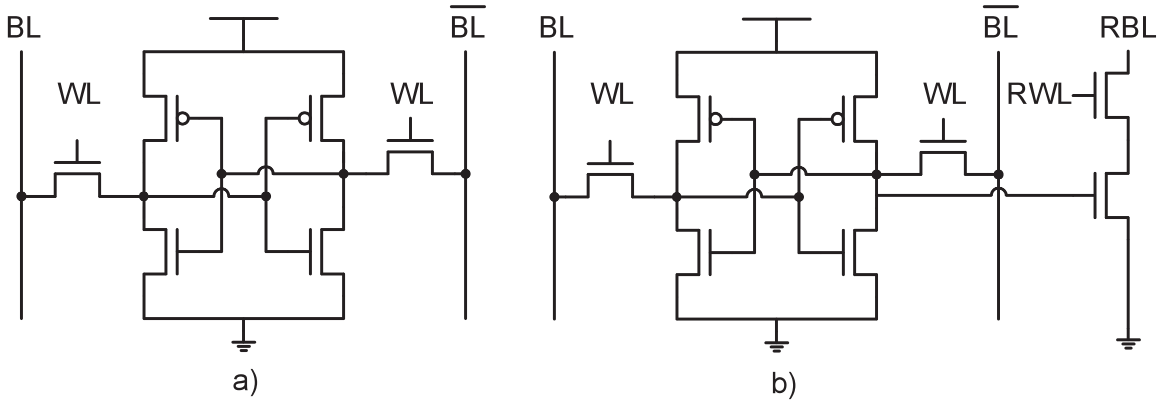


Figure 4.3: The candidate bitcells: a) Differential 6T b) 8T.

setup and importance sampling methodology; Section 4.4 presents our results, and Section 4.5 concludes this chapter.

4.2 Candidate SRAM Architectures

4.2.1 Differential 6T Bitcell

For the differential 6T bitcell shown in Figure 4.3*a*, a read is performed by precharging and floating the bitlines (BL and $\bar{BL}$) in the desired columns at V_{DD} and asserting the wordline (WL) in the desired rows. The bitcell pulls down either BL or $\bar{BL}$ depending on the bitcell's state, and the voltage differential is detected using a sense amplifier. A write is performed by driving opposite values onto the bitlines and asserting WL, overwriting the value held in the bitcell.

Bitcells are susceptible to four prominent failure modes: read upset, write, timing and hold. During a read operation on a 6T bitcell, noise is injected from the bitline through the pass gate transistors to the node holding a ZERO value. Read upset occurs when the voltage transient on the ZERO node causes the bitcell value to flip. Read upset tolerance depends heavily on the cell ratio (on-current ratio of the pull down to pass gate transistors) as well as the feedback from the cross coupled inverters. Write stability requires adequate pass gate transistor strength to overwrite the value held in the bitcell. The most critical transistors for a write are the pass gate device connected to the BL at a ZERO value and the pull up PMOS holding the bitcell node to ONE. The requirements for both read and write stability place contradicting requirements on pass gate strength. For this reason, at lower V$_{DD}$s, 6T bitcells must be sized up substantially or doped differently to achieve both read and write stability, or they may not be able to achieve both.

4.2.2 SRAM Assist Circuits

As an alternative to sizing the 6T bitcell, assist circuits can be used to prevent failure. Read upset can be prevented by lowering the WL voltage in relation to the SRAM array V_{DD} [43]. This reduces the cell ratio of the bitcell but increases delay, hurts write stability and requires an additional voltage source. To prevent write failures a dual-V_{DD} WL or additional write assist circuitry can be employed. In a dual-V_{DD} scheme, WL voltage is only reduced when a read access is performed. This requires additional decoding and a more complex wordline driver to select between two WL voltages.

Dual-V_{DD} WL and other schemes have also been proposed to enhance write robustness. During a write operation, the WL voltage can be increased above the SRAM array V_{DD}, increasing the effective pass gate strength [44]. Another way to increase the pass gate strength and increase write stability is to pull the bitline to a negative voltage [45]. The negative BL voltage must not turn on unaccessed devices on the same BL and must not cause intolerable junction leakage. Both the dual-V_{DD} and negative BL techniques require an additional voltage source. Another write assist method droops the SRAM array V_{DD} and GND during a write [47][46]. This reduces the strength of the cross coupled inverters that hold the bitcell state, facilitating write. V$_{DD}$ drooping can be implemented with diode drops in shared headers and footers. The drooped supplies must be shared in a row or column, and unaccessed drooped bitcells must retain their state.

4.2.3 8T Bitcell

The 8T in Figure 4.3*b* uses two additional transistors over the differential 6T bitcell to isolate the read and write paths [42]. This enables separate optimization of the read and write mechanisms. The two stacked NFETs are connected to additional read word and read bit lines (RWL and RBL) as well as one bitcell node to perform a single-ended read. This read circuitry eliminates the read upset failure mode. A write operation is performed similarly to a write in the differential 6T bitcell, however, since the pass gate devices and cross coupled inverters are not used for reading, they can be optimized solely for write.

The 8T bitcell has the same timing failure mode as the differential 6T. However, since the 8T read is single ended, differential sense amplifiers cannot be used to improve delay and minimize read bitline swing. For our study, we sense an 8T read using the same sense amplifier structure with one input tied to a reference voltage. The reference voltage must be sufficiently below V_{DD} to sense the read of a ONE. This necessitates that the bitline falls below the reference voltage to sense a ZERO, increasing the delay and bitline swing for an 8T read. In our study, this delay must be recuperated by optimizing the stacked NFETs used for reading. Increasing the strength of those devices does not exacerbate other failure modes, however, it incurs area and energy penalties.

4.3 Reliability Analysis using Importance Sampling

4.3.1 Scaling Methodology for Iso-robustness Low V_{DD} Operation

When SRAM bitcells are naively scaled into the near-V$_{TH}$ V_{DD} region, significant energy gains are achieved, but random dopant fluctuations (RDF) and other process variations lead to functional failures and low yield. In our study we will examine the robustness of 6T and 8T SRAM in a 65nm process when V_{DD} is scaled to the near-V$_{TH}$ region. We constrain all bitcells at all V_{DD}s to have the same robustness as the differential 6T bitcell at 1V with sizes taken from commercial designs. To meet these constraints as the bitcells are scaled into the near-V$_{TH}$ region, the bitcell device's geometry and dopings are optimized, or assist circuits are tuned. The delay, density, and energy of the final bitcells are compared to find the advantages and disadvantages of all designs.

4.3.2 Sizing and Doping Methodology

For our sizing and doping study, we will adjust device strengths to prevent functional failure when V_{DD} is reduced. We constrain the bitcell delay to scale with logic, such that memory latency (in cycles) is not affected when V_{DD} is scaled. The wordline driver, bitline driver, and bitcell delays are monitored in this study. The bitcell delay for a read is measured to the time when adequate bitline swing is developed for differential or single-ended sensing with a commercial current-mode sense amplifier.

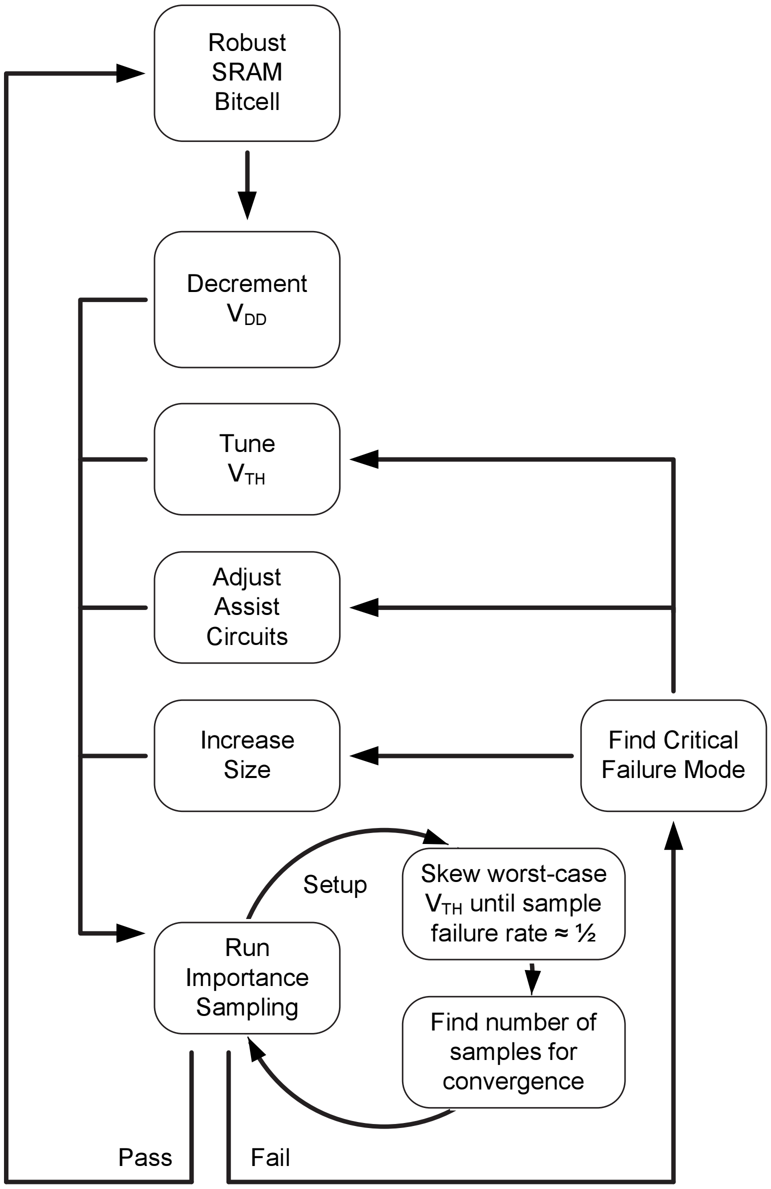


Figure 4.4: Design Methodology for yield-driven near-threshold SRAM.

In modern SRAM designs, V_{TH} is optimized separately from V_{TH} for logic to improve robustness and performance. The 65nm process used in this study has a nominal V_{DD} of 1.1V and uses separate NFET V_{TH}s of 560mV and 520mV for the pass gate and pull down devices, respectively. These V_{TH}s are carefully chosen by manufacturers to enhance performance at nominal V_{DD}, however as V_{DD} is scaled down, the criticality of failure modes and thus the optimal V_{TH}s change. Optimizing V_{TH} can help SRAM meet delay requirements as well as control the current ratios between devices to balance probabilities of different failure modes. When V_{TH} is tuned, σV_{TH} is calculated appropriately according to the device models. Circuit designers have limited flexibility to tune V_{TH}, therefore, in this study we will optimize the device geometry alone and also geometry with individual device V_{TH}s. Reasonable limits are placed on V_{TH} to ensure realistic doping concentrations and tolerable leakage power.

4.3.3 Assist Circuit Methodology

In our study of assist circuits, we will maintain bitcell robustness as V_{DD} is scaled down by adjusting the peripheral circuits. Assist circuits are unnecessary for the 8T bitcell because there is no read upset failure mode and write stability can be maintained with minimal sizing. The 6T bitcell design in our study is taken from a commercial design optimized for super-V$_{TH}$ operation and no device sizing or V_{TH} tuning is performed. To maintain read stability, a dual-V_{DD} WL with reduced read voltage is used. This read assist circuit incurs a delay penalty, precluding iso-latency voltage scaling, so there is no delay constraint for the assist circuit study. For write robustness, three methods will be compared: overdriven WL, negative BL, and supply rail drooping. In the latter two cases, the assist circuits must be adjusted appropriately to not disturb unaccessed bitcells. The resulting decrease in bitcell performance and changes in energy consumption are measured.

4.3.4 Robustness Calculation using Importance Sampling

At this point, an accurate metric of SRAM robustness is necessary to determine when optimization is complete. SRAM robustness is often measured using SNM because it is relatively easy to compute. However, SNM does not consider the dynamic nature of noise injection into bitcells. Since the probability of injecting the same amount of noise changes as V_{DD} is scaled, SNM does not translate directly to SRAM yield. Corner cases can also be used to measure robustness, however, in general the supplied corner cases only consider global variation and not device mismatch. Since mismatch has a strong effect on SRAM yield, these cases are not sufficient. Corner cases involving mismatch can be performed but have several drawbacks. First, different transistors have differing criticality for SRAM functionality, but in corner case analysis the same amount of variation is placed on each device, making the analysis incomplete. Second, calculating SRAM yield based on corner case simulations is non-trivial.

For a complete look at SRAM reliability sampling methods like Monte Carlo are necessary. In Monte Carlo sampling, the number of passing bitcells is divided by the total number of iterations (n) to find the expected yield, as shown in Equation 4.1 [53] [54] [55]. Process parameters such as V_{TH} and gate length are selected from a probability density function (PDF), which represents the natural variation in the process parameter. As shown in Figure 4.5, the PDF of V_{TH} in SRAM devices is modeled as a normal distribution. Since caches contain many bitcells, the failure rate of each one must be very low in order to have high yield for the cache. For example, to have a 99% yield for a small 8kB SRAM, the bitcell failure rate must be 1.53×10^{-7}. To calculate this bitcell yield using Monte Carlo, at least 10 million simulations must be performed, making this procedure computationally intensive. For larger caches, the required bitcell failure rate is even lower and complete Monte Carlo analysis is almost infeasible.

$$Y = \frac{1}{n}\sum_{n} f(x)\ where\ f(x) = 1, pass; 0 fail \tag{4.1}$$

$$Y = \frac{1}{n}\sum_{g(x)} \frac{p(x)}{g(x)} \tag{4.2}$$

$$p(x) = \prod \frac{1}{\sigma\sqrt{2\pi}} exp \frac{(x-\mu)^2}{2\sigma^2} f(x) \tag{4.3}$$

$$g(x) = \prod \frac{1}{\sigma\sqrt{2\pi}} exp \frac{(x-\mu+4\sigma)^2}{2\sigma^2} \tag{4.4}$$

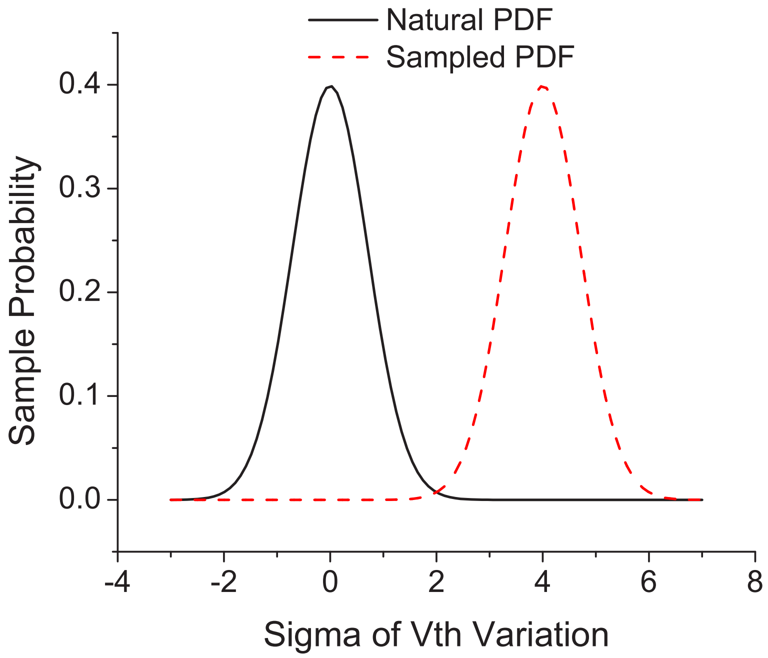


Figure 4.5: For importance sampling, device V_{TH}s are selected from the sample probability density function (SPDF). The SPDF is created by shifting the natural V_{TH} PDFs into the failure region by introducing mismatch between devices. By selecting V_{TH} values from this region of interest, the SRAM failure modes can be more quickly analyzed.

$$Y = \frac{1}{n} \sum_{g(x)} \frac{\prod \frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{(x-\mu)^2}{2\sigma^2}} f(x)}{\prod \frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{(x-\mu+4\sigma)^2}{2\sigma^2}}} \tag{4.5}$$

For our study, we choose importance sampling as an efficient and accurate way of calculating SRAM robustness. As shown in Figure 4.5 and Equations 4.2 to 4.5, the importance sampling technique chooses a new sampling PDF (SPDF) for each transistor so that more failures are simulated. The V_{TH} of each transistor is shifted by the value sampled from the PDF plus 4σ, to be justified below, to introduce enough mismatch into the bitcell to increase the probability of failure. Since the natural occurrence of these highly skewed devices is rare, the importance samples are then weighted by the ratio of the probability of the large V_{TH} shift in each transistor to the probability that these V_{TH} shifts were sampled. These weighted values are then used to calculate the bitcell yield. This method allows us to accurately measure the region of interest where SRAM can fail with greatly reduced computational complexity.

Since the sampling PDF and number of importance samples have a large impact on experimen-

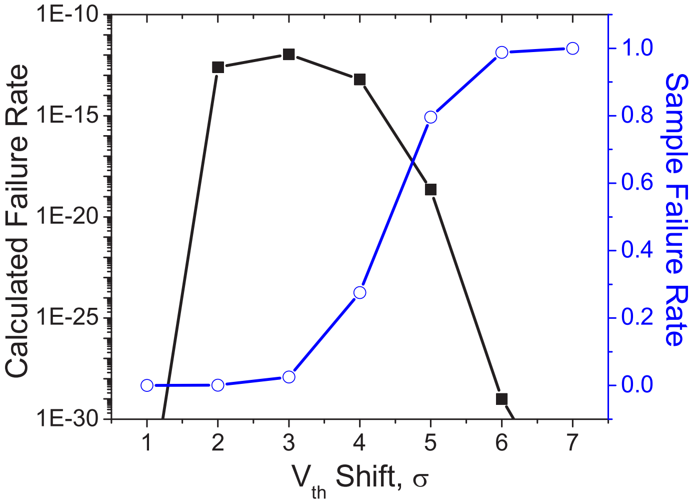


Figure 4.6: The natural probability density function (PDF) is shifted into the failure region to create the sample PDF (SPDF). If the SPDF is too similar to the PDF, then simulation runtime to calculate yield is long because few failures are seen. If the SPDF is too different from the PDF then more samples are necessary for accurate yield calculations.

tal results, they were carefully chosen to maintain accuracy in the simulation while still reducing simulation runtime. A small V_{TH} shift in the sample PDF would not introduce a large number of failures, thus negating the variance reduction effect of importance sampling. Conversely, an overly large V_{TH} shift introduces failures, but causes the sample weighting to be small and reduces the accuracy of the simulation. A differential 6T bitcell is studied to find the optimal sampling PDF. As shown in Figure 4.6, with less than a 4σ V_{TH} shift the sample failure rate is very low. Above a 4σ V_{TH} shift the calculated failure rate drops and is inaccurate. Therefore, a 4σ shift is chosen for our study. After a sufficient number of importance samples have been performed, the calculated failure rate converges to its final value. Figure 4.7 shows the calculated failure rate converging and more samples are taken. We determine that 20,000 samples are sufficient for accurate results. To measure the failure rates in our study with Monte Carlo, at least 10^{12} samples are needed, making importance sampling 50 million times faster.

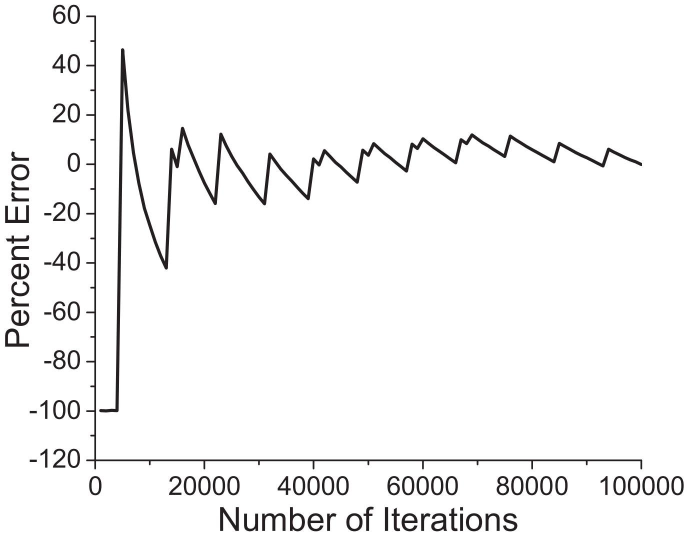


Figure 4.7: After a sufficient number of importance samples have been simulated, the calculated yield will converge to the correct value.

4.4 Experimental Results

4.4.1 Bitcell Sizing and Doping in Near-threshold SRAM

We examine the area and energy of 6T and 8T bitcells when V_{DD} is scaled down, robustness is maintained through sizing, and delay is constrained to scale with logic. The analysis is performed with and without the ability to individually adjust the pass gate, pull down and pull up device V_{TH}s. The iso-robustness bitcell sizings are plotted in Figure 4.8. V_{TH} tuning dramatically reduces the required bitcell area at low voltage. V_{TH} is often set higher for pass gate devices than for pull down devices to prevent read upset failures at super-V$_{TH}$ V_{DD}. However, when V_{DD} is scaled to near V$_{TH}$ V$_{DD}$s, this V_{TH} selection makes the pass gates too weak for write robustness. This effect is especially strong when the pass gates enter sub-V$_{TH}$ operation, but other devices are still in the near-V$_{TH}$ regime. If V_{TH} is a fixed parameter, set an a value optimized for super-V$_{TH}$ operation, then SRAM bitcells must be sized by 400% at 500mV, and voltage scaling to the sub-V$_{TH}$ regions is not practical for iso-robustness operation. Tuning V_{TH} enables balancing of the SRAM failure modes. In our study, the highest density, robust SRAM is achieved by increasing pass gate V_{TH} to prevent read upset for near-V$_{TH}$ SRAM. As V_{DD} is further scaled, the optimal

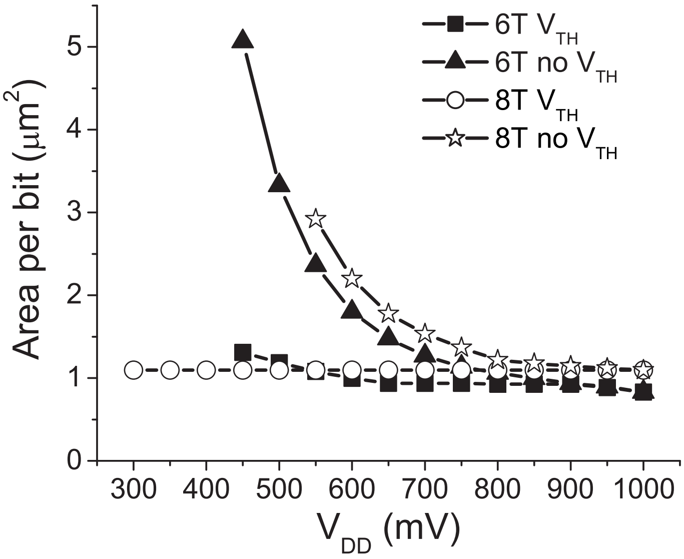


Figure 4.8: Bitcells can be sized up to maintain robustness when V_{DD} is scaled. The density of iso-robustness sub-V$_{TH}$ SRAM is improved when V_{TH} tuning is used.

pass gate V_{TH} is lower because write failures become critical. When V_{TH} is tuned, robustness can be maintained in 6T SRAM at 500mV with a 40% area penalty. Across all V$_{DD}$s studied, V_{TH} tuning with minimal sizing is sufficient to maintain robustness in 8T SRAM, enabling high-density low-voltage memory.

V_{DD} scaling from 1V to 500mV reduces dynamic energy by more than 50% for all bitcells studied, with a 61% reduction for 8T SRAM with V_{TH} tuning, as seen in Figures 4.9 and 4.10. In our study we consider energy from the wordline drivers, bitline drivers and bitcells only. If other memory peripheries, such as the decoder and sense amplifiers, are voltage scaled with the SRAM bitcells, then greater energy gains than those reported are possible. Without V_{TH} tuning and at low V$_{DD}$s, bitcells must be aggressively sized to control relative device strengths under RDF variation and maintain robustness. Device sizing substantially increases wordline and bitline capacitances, thus reducing the energy benefit of voltage scaling. For sub-V$_{TH}$ robustness without V_{TH} tuning, devices must be sized up to a level where the energy benefit is eliminated. When V_{TH} tuning is used, less dramatic sizing is needed, keeping capacitance and energy lower. Using V_{TH} tuning for near-V$_{TH}$ 550mV SRAM reduces dynamic energy by 44% and 56% for 6T and 8T SRAM, respectively, over the fixed V_{TH} case. For 8T SRAM, iso-robustness operation at 300mV

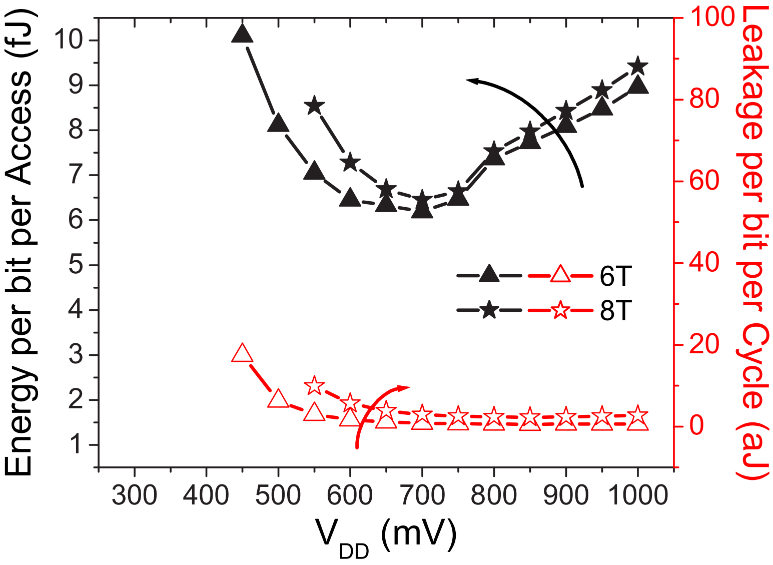


Figure 4.9: When bitcells are sized for robustness without V_{TH} tuning, the energy benefits from iso-robustness scaling of 65nm SRAM to the near-V$_{TH}$ region are limited because wordline and bitline capacitances from upsized devices are prohibitive.

is obtained with little device sizing and an 83% energy reduction is achieved.

Above 500mV, leakage energy per cycle is relatively constant, whereas below 500mV, leakage increases dramatically. Although leakage power scales down linearly with V_{DD}, leakage energy per cycle is also proportional to delay, which increases exponentially in the sub-V$_{TH}$ region. Since dynamic energy decreases and leakage increases when V_{DD} is scaled down, a minimum energy point (E_{MIN}) is achieved at some intermediate voltage (V_{MIN}) [37]. As seen in Figures 4.11 and 4.12, V_{MIN} and E_{MIN} are heavily dependent on activity factor, which we define as the average fraction of bitcells accessed per cycle. For L1 caches, which are generally small with high activity, total energy is almost entirely dynamic, making voltage scaling a desirable method for energy reduction. Based on typical memory access patterns, an 8T 8-way 1kB L1 cache could have an activity factor of 10^{-2} and a V_{MIN} of 450mV. In L2 caches, which are larger with lower activity, the benefits of voltage scaling are reduced and V_{MIN} rises. A large L2 cache could easily have an activity factor lower than 10^{-6}, making voltage scaling below 850mV detrimental.

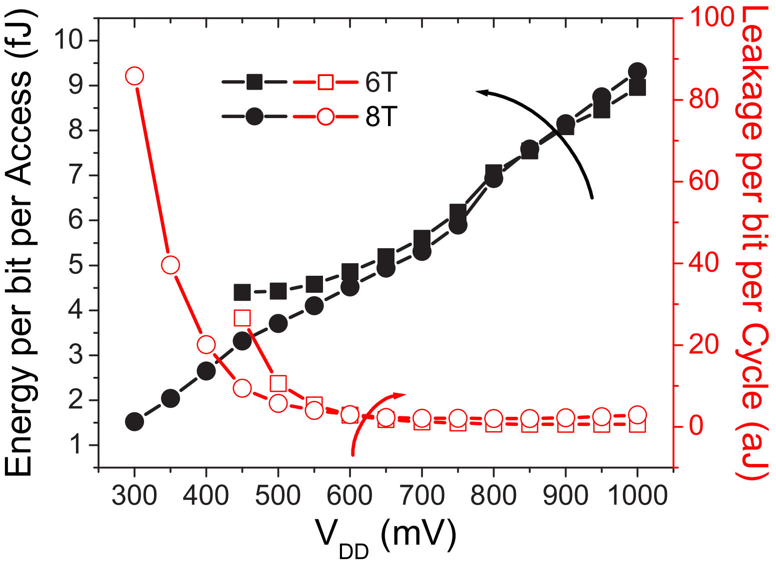


Figure 4.10: When bitcells are sized and V_{TH} is tuned for robustness, dynamic energy can be reduced by as much as 83% in iso-robustness SRAM using voltage scaling.

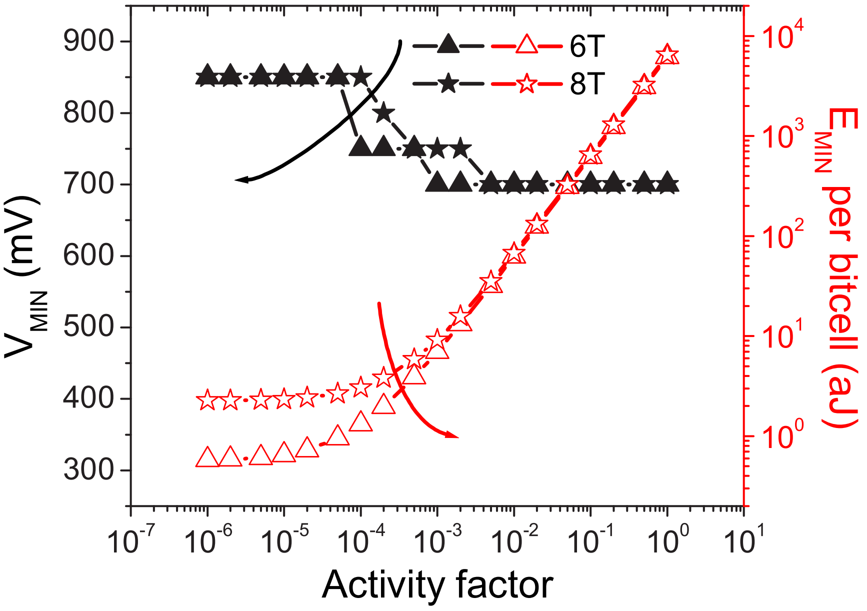


Figure 4.11: When bitcells are sized for robustness without V_{TH} tuning, the V_{DD} for minimum energy computing is above 700mV, because the large device sizes needed to maintain robustness in near-V_{TH} SRAM result in large capacitances and switching energy.

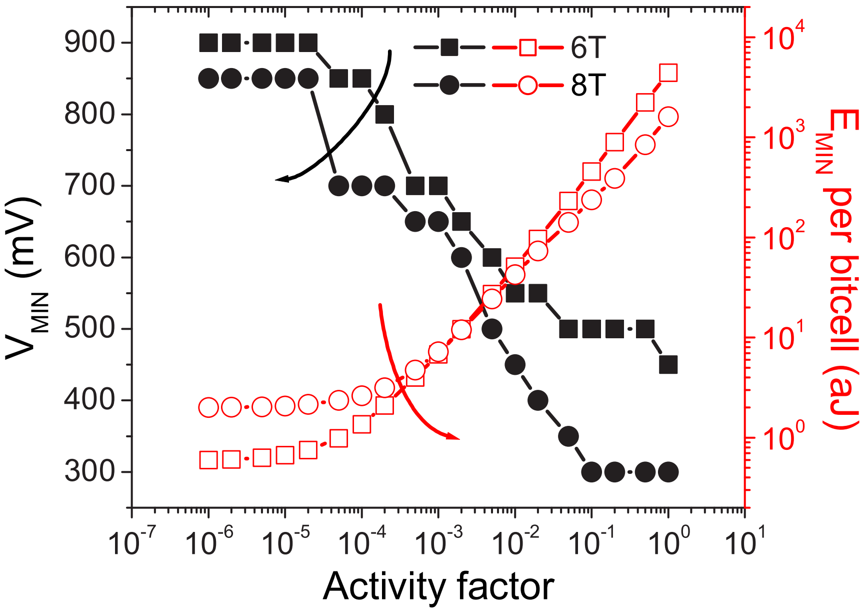


Figure 4.12: When bitcells are sized and V_{TH} is tuned for robustness, 8T L1 caches with high activity factor can benefit from voltage scaling to 300mV.

4.4.2 Assist Circuits for Near-threshold SRAM

Assist circuits can be used to increase SRAM robustness and enable near-V_{TH} operation. The dual-V_{DD} WL, negative BL, and supply drooping assist circuits considered in this study increase SRAM stability by modifying control or V_{DD}s during accesses. The voltage levels necessary for iso-robustness operation are shown in Figure 4.13. Of the three write assist circuits studied, only overdriven WL enables sub-V_{TH} SRAM. A functional minimum-sized SRAM cell with no V_{TH} tuning and an SRAM array V_{DD} of 300mV, requires a write WL voltage of 650mV. This near-V_{TH} WL voltage requires additional access energy and precludes unaccessed bitcells on the WL. Negative BL and supply drooping can keep robustness high when V_{DD} is scaled to 650mV and 600mV respectively. Below these V_{DD}s, the aggressive assist circuits needed to maintain robustness disturb unaccessed bitcells. For the negative BL scheme, when the BL is driven below GND for a write, pass gates of unaccessed bitcells are partially on and can cause erroneous writes. For unaccessed bitcells with supply drooping, process variation and supply transients cause the loss of state.

An underdriven WL, helps to prevent read upset failures but also reduces performance, shown

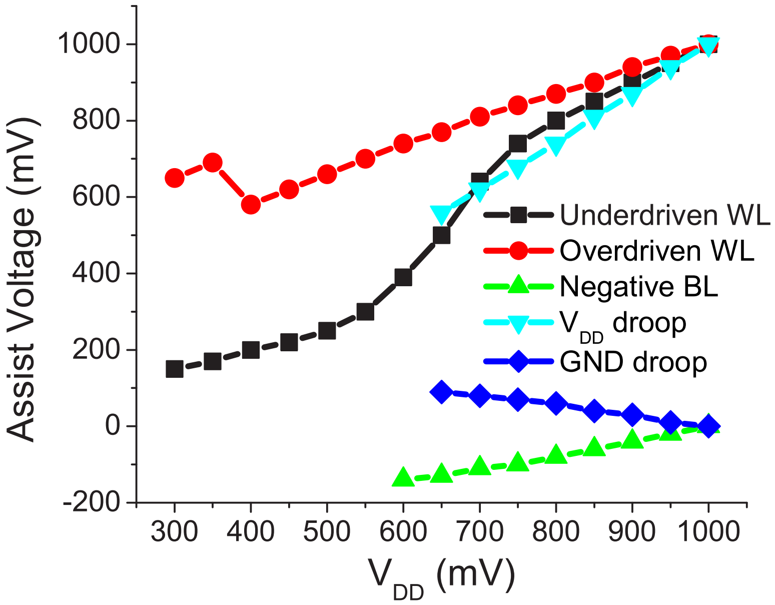


Figure 4.13: Assist circuits modulate SRAM voltages and can maintain robustness as V_{DD} is scaled down. Underdriving the WL during read prevents read upset. Overdriving the WL during write can enable iso-robustness sub-V_{TH} SRAM. Negative BL and supply drooping disturb unaccessed bitcells below 600mV.

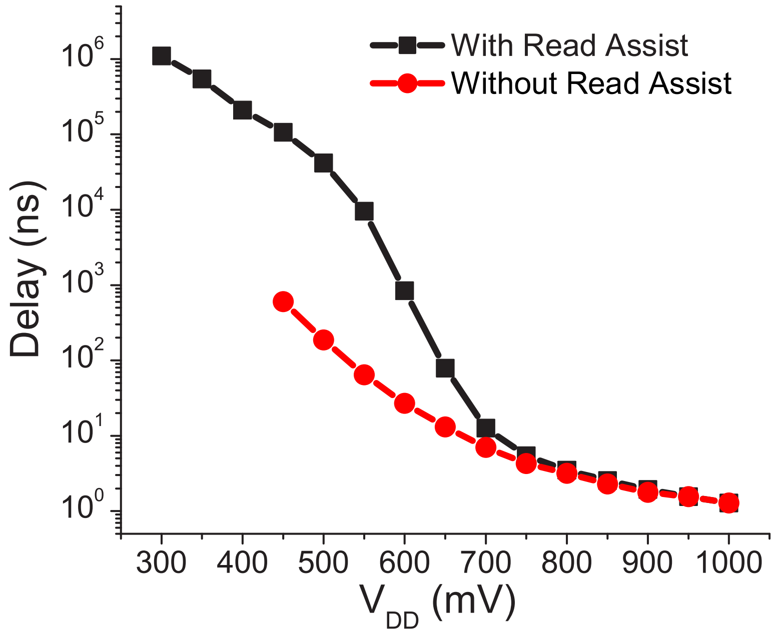


Figure 4.14: The delay of iso-robustness 6T bitcells is significantly greater with read assist than with sizing and doping.

in Figure 4.14. At 500mV, the WL voltage must be reduced to 250mV to have the same robustness as the unassisted bitcell with a 1V supply, resulting in a 200x increase in bitcell delay. This excessively large delay also manifests itself as intolerable leakage energy, shown in Figure 4.15. As a result, when assisted SRAM circuits are scaled to the near-V$_{TH}$ region, leakage energy dominates and the V_{MIN} never falls below 600mV, regardless of activity factor. The active energy for the three write assist circuits is almost the same, and the V_{MIN} for all assist circuits is shown in Figure 4.16.

4.5 Conclusions

We have compared 6T and 8T bitcells in various voltage domains with an iso-robustness condition. Our study is enabled by using importance sampling to accurately calculate SRAM yield 50 million times faster than with Monte Carlo sampling. We find that energy gains of 50% can be achieved for small caches by halving V_{DD} to 500mV with no decrease in robustness and a small area overhead. At 300mV, 8T SRAM with low V_{TH} devices can deliver an 83% energy reduction

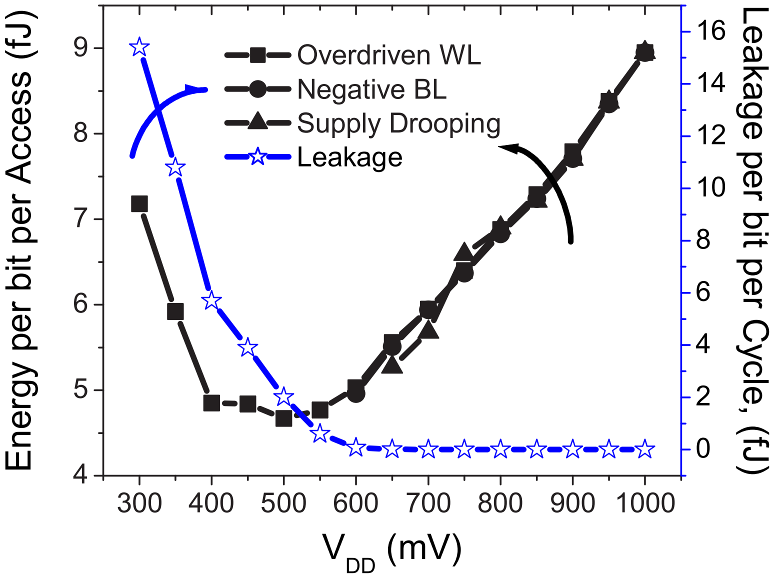


Figure 4.15: When assist circuits are used to maintain SRAM robustness, dynamic energy can be reduced by 50% by halving V_{DD}.

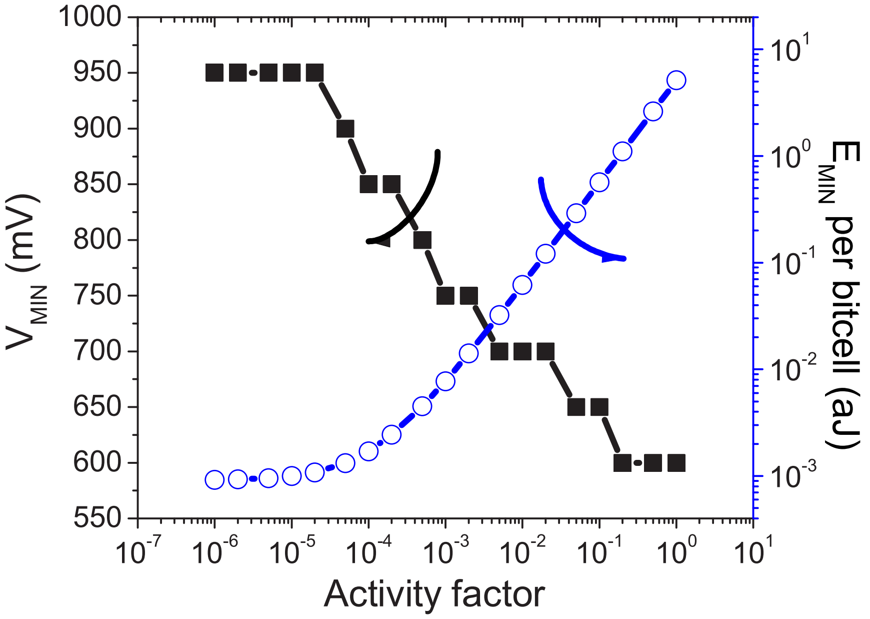


Figure 4.16: When assist circuits are used to maintain SRAM robustness, V_{MIN} for caches with high activity factor can be as low as 600mV.

over the nominal case. For L1 caches, the V$_{DD}$ for minimum energy iso-robustness operation can be as low as 300mV, making voltage scaling a desirable technique for low-energy computing. Assist circuits can only enable iso-robustness SRAM to scale to 600mV before delay and leakage become prohibitive. The method shown in this paper assesses design tradeoffs in SRAM quickly and accurately, allowing a designer to select an appropriate SRAM architecture and sizing.

CHAPTER 5

Crosshairs SRAM - An Adaptive Memory for Mitigating Parametric Failures

5.1 Introduction

In Chapter 4 we examined robust, low V$_{DD}$, traditional SRAM designs for cubic-millimeter microsystems. Sizing, V$_{TH}$ selection, and assist circuits were used to maintain stability margins as V$_{DD}$ was scaled to reduce power. Each of the techniques used to enable low V$_{DD}$ had tradeoffs in area, leakage, or complexity, which were quantitatively analyzed. In this chapter and Chapter 6, we present novel bitcell techniques to enable low-power low V$_{DD}$ SRAM with improved stability margins and more favorable design tradeoffs.

Crosshairs mitigates the effects of process variations, which are exacerbated at low V$_{DD}$. Excessive process variation causes parametric failures (PFs), including timing and stability failures. For example, voltage scaling amplifies the effects of V$_{TH}$ mismatch caused by random dopant fluctuations [56] and lithographic double patterning induced gate length variation and V$_{TH}$ rolloff [57]. As a result, SRAM requires higher levels of error correction coding (ECC) [58] and redundancy [59] to satisfy yield requirements. We propose the Crosshairs method to detect and adaptively correct PFs. Crosshairs tunes the SRAM's power and ground supply networks to mitigate excessive variation. It improves yield with respect to timing and stability constraints. The Crosshairs bitcell has the same area, transistor count, and number of metal layers as a commercial design.

5.2 Crosshairs SRAM Method

5.2.1 Controlling Bitcell Power Supplies

To identify PFs, a BIST performs March tests on the SRAM. When it detects a failure, it determines the nature of mismatch in the bitcell by checking if write-ZERO/read-ZERO or write-ONE/read-ONE accesses failed. The BIST then uses this information to tune V_{DD} and GND of each bitcell inverter with respect to its cross-coupled counterpart, canceling process variation and restoring bitcell functionality. Each bitcell has connections to left and right vertical power rails (V_{DDL} and V_{DDR}) and horizontal ground rails (GND$_L$ and GND$_R$). It is identical to a commercial differential 6T design except that the vertical V_{DD} rail is split into V_{DDL} and V_{DDR} (Fig. 5.1). This modification does not require a larger bitcell or more metal layers (Fig. 5.2). Crosshairs eliminates PFs by adjusting V_{DD} in the column and GND in the row where a PF occurs (Fig. 5.3). Thus, the orthogonal tuned supply rails target PFs at their intersection. Each column shares V_{DDR} and V_{DDL} rails and adjacent rows share GND$_R$ and GND$_L$ rails (Fig. 5.4). To tune V_{DD}, PMOS headers connect V_{DDR} and V_{DDL} to one of two global power supplies (VDD_{HI} and VDD_{LO}). Similarly, NMOS footers connect GND$_R$ and GND$_L$ to either GND$_{HI}$ or GND$_{LO}$ (Fig. 5.5). The BIST generates control signals for the headers and footers. An on-chip linear regulator can generate the global V_{DD}s and GNDs. The voltage difference between these global supplies is defined as the Crosshairs tuning voltage.

5.2.2 Fixing Parametric Failures

Tuning the supplies of each bitcell inverter with respect to its cross-coupled counterpart cancels process variation and eliminates PFs. Initially the stronger V_{DD} and GND (VDD_{HI} and GND$_{LO}$) supply all bitcell inverters. When writing a ZERO to node D in Fig. 5.1, the left pass gate (PG$_L$) overpowers the left pull up (PU$_L$), pulling D low enough to initiate the write mechanism. Process variation can create a write PF by making PG$_L$ too weak with respect to PU$_L$. To increase write margin and fix this PF, Crosshairs weakens PU$_L$ by connecting V_{DDL} to VDD_{LO} and GND$_R$ to GND$_{HI}$, reducing the likelihood of a write PF by 9x based on importance sampling Monte Carlo SPICE simulations [5]. When reading a ZERO from node D, charge from the bitline (BL) is

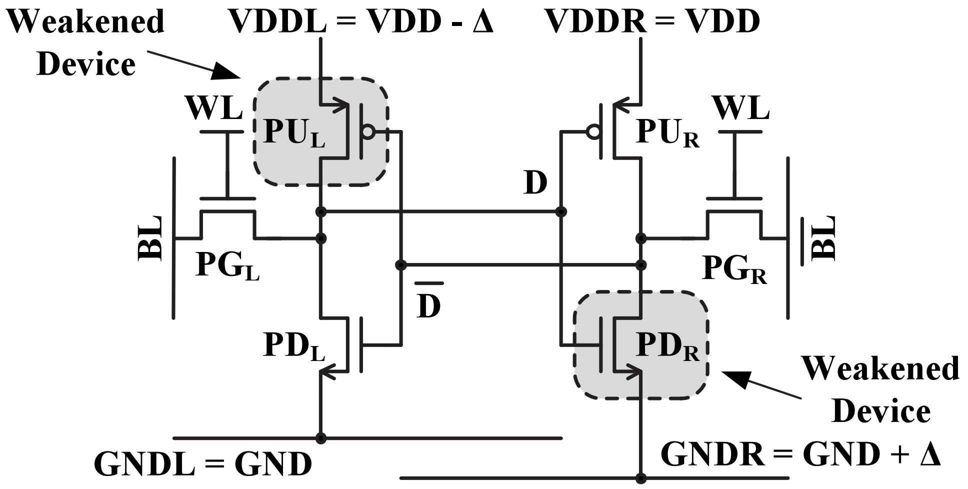


Figure 5.1: Crosshairs recovers parametric failures (PFs) by separately tuning the V_{DD} and GND supplies of each inverter within a bitcell.

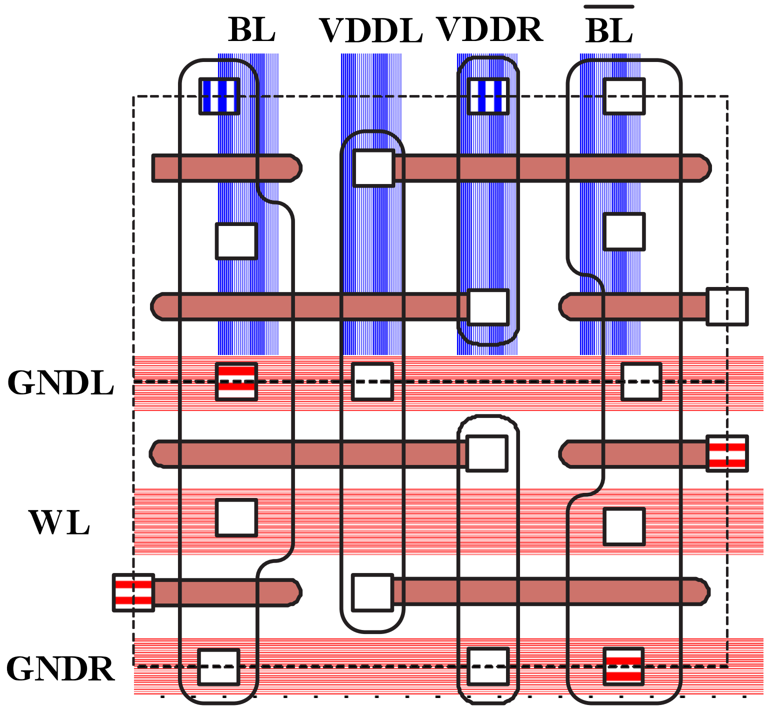


Figure 5.2: The Crosshairs bitcell is a minimally modified commercial differential 6T design that does not require larger area or more metal layers.

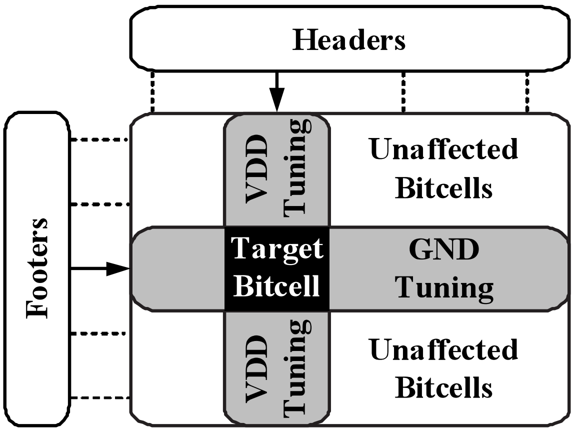


Figure 5.3: Crosshairs shares V_{DD} within a column and GND within a row. The orthogonal, tuned supplies target failing bitcells at their intersection.

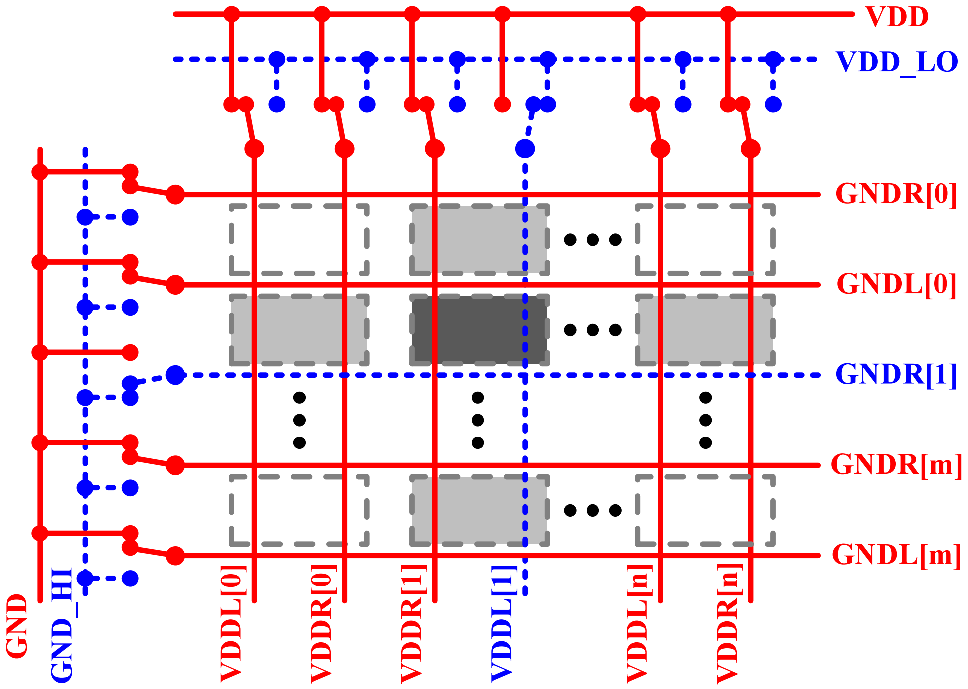


Figure 5.4: Header and footer cells control V_{DD} and GND potentials, increasing array efficiency.

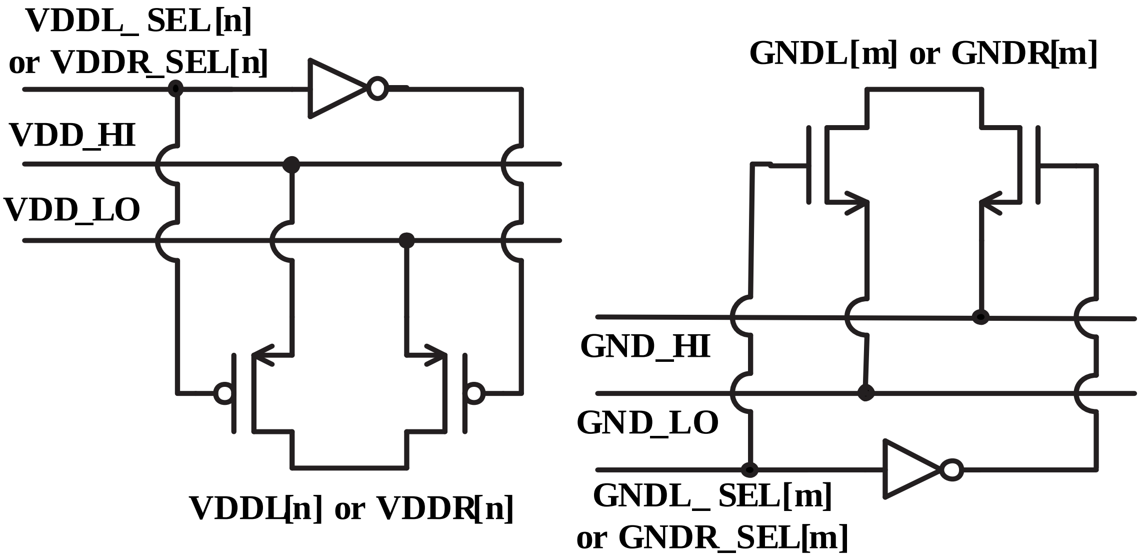


Figure 5.5: Headers connect each V_{DD} column (V_{DDL}, V_{DDR}) to one of two global supplies (VDD_{HI}, VDD_{LO}). Similarly footers connect each GND row (GND$_L$, GND$_R$) to global grounds (GND$_{HI}$, GND$_{LO}$).

injected onto D, potentially causing a read upset PF by overwriting the value to a ONE. This is more probable if process variation causes the left pull down (PD$_L$) to be too weak with respect to PG$_L$ or the timing constraint. To correct this, Crosshairs weakens the right pull down (PD$_R$) device by connecting GND$_R$ to GND$_{HI}$. Similarly the left pull up (PU$_L$) device is weakened by connecting V_{DDL} to VDD_{LO}. In this configuration the bitcell holds a stronger ZERO and the probability of a read PF decreases by 3x, based on importance sampling [5]. A larger simulated static noise margin (SNM) reflects the increase in read stability (Fig. 5.6a). As seen from the previous examples of write and read PFs, Crosshairs uses the same voltage configuration to improve both read-ZERO and write-ZERO margins. Similarly, lowering V_{DDR} and raising GND$_L$ increases read-ONE and write-ONE margins. Thus, to properly apply the Crosshairs algorithm, the BIST does not need to determine whether the write or read access failed. Rather, it must determine only whether a ONE or ZERO access failed. This allows the BIST to gather all the information it needs about process variation in the SRAM array using simple March test algorithms.

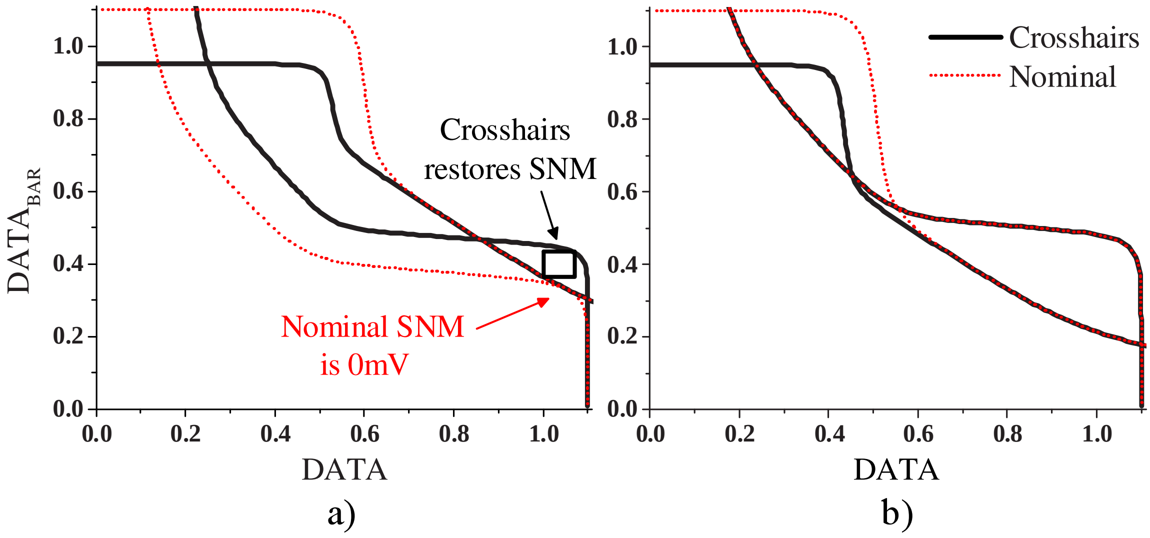


Figure 5.6: a) Crosshairs restores simulated read SNM in a PF cell. b) It has less impact on non-PF cells in the same column or row as a PF cell because either V_{DD} or GND is tuned but not both.

5.2.3 Net Reduction in SRAM Failures

Crosshairs tuning can positively or negatively impact non-PF cells in the same column or row as a PF. It adjusts the V_{DD} rails in each column based on the process variation in the PF cell, which does not necessarily reflect the variation in the other cells. Similarly, Crosshairs will impact the GNDs of cells in the same row as a PF cell. However, Crosshairs will not tune both V_{DD} and GND in non-PF cells, greatly decreasing the potential negative impact on stability margins (Fig. 5.6b). The distribution of each bitcell with process parameters is weighted at the mean, with few cells at tails of the distribution that fail yield criteria (Fig. 5.7a). Applying Crosshairs tuning relaxes the yield criterion on one end of the distribution, but tightens it at the other tail. The conditional PDFs for PF and non-PF cells dictate that Crosshairs is likely to fix a given PF, but unlikely to cause new errors. This probability is calculated using importance sampling as 95.5% for a 128x256 array with a 20mV tuning voltage [5] (Fig. 5.7b).

5.2.4 Header and Footer Sizing

Headers and footers require proper sizing to prevent IR drop from impacting robustness but should be small for a low area overhead. Fig. 5.8 shows a simulation demonstrating the affect of these sizes on robustness. The plot shows robustness in terms of the maximum V_{TH} mismatch that

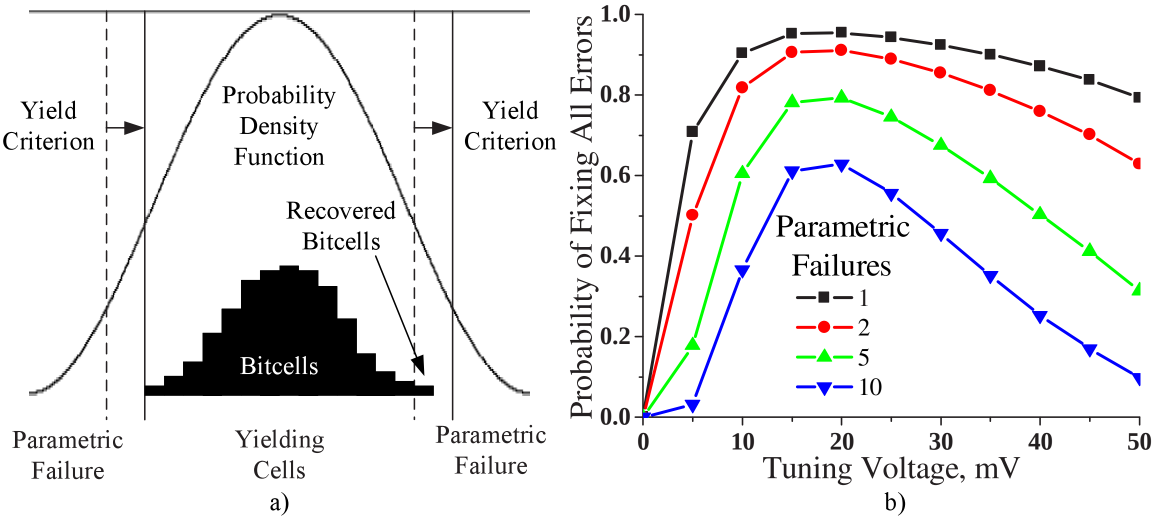


Figure 5.7: Crosshairs has a high probability of fixing PFs and a low probability of creating PFs, resulting in a net reduction in simulated failures.

the bitcell can tolerate without functional failure. For each V_{DD} column, we select a header width of 2x the bitcell PU device. Further increasing header size achieves only modest improvements in stability. Crosshairs requires only a small header since only one accessed cell per column draws current from the V_{DD} rail. However, every bit can simultaneously draw current from the same GND rail. As such, the footer size is 2x the total PD width for one word. The presented array uses a 128-bit word and footer size decreases proportionally with word length.

5.3 Measurement Results

We fabricated and measured 70 chips with 128x256 32kb Crosshairs SRAM banks in a 45nm CMOS process (Fig. 5.9). We designed Crosshairs with feedback from the foundry to violate logic design rules, as is typical for SRAM. This allows the bitcell to match the area of a commercial differential 6T SRAM design. Crosshairs decreases array efficiency by 12.5% because of additional peripheral circuits.

5.3.1 Recovering Timing Failures

Local process variation creates slow bitcells, which then dictate the overall performance of an SRAM array. By targeting slow bitcells, Crosshairs mitigates process variation and increases array

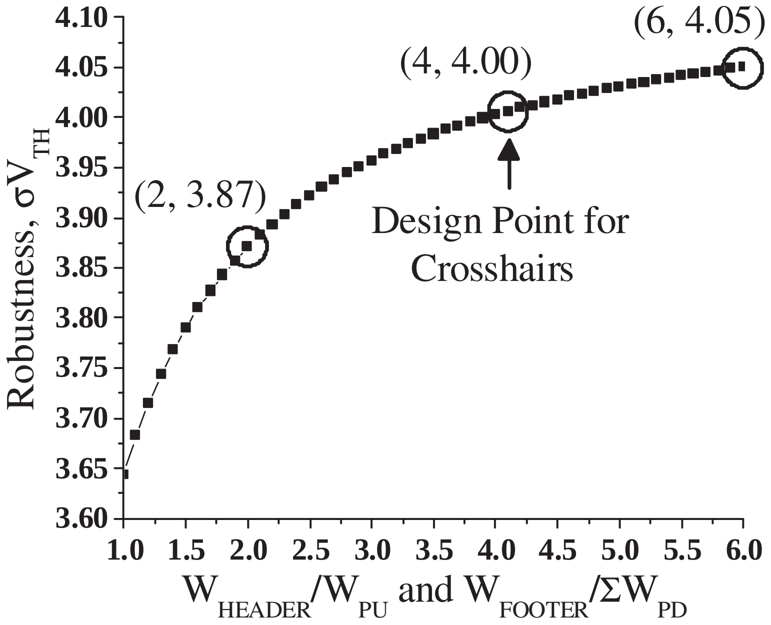


Figure 5.8: Headers and footers are optimally sized to create stable virtual supply rails without excess area overhead, based on simulations.

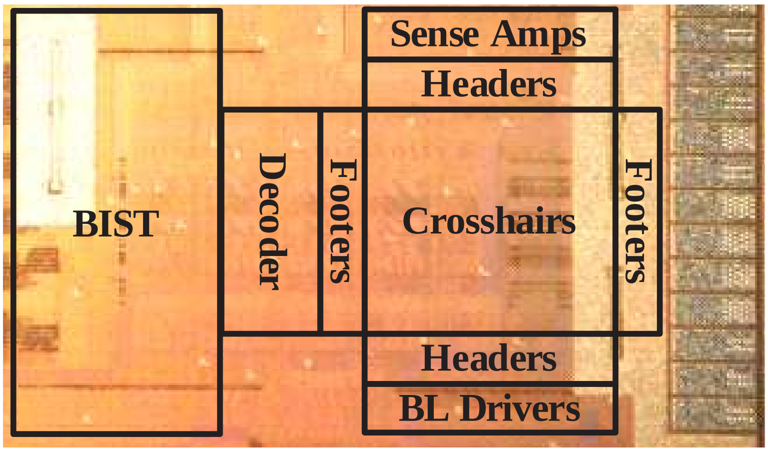


Figure 5.9: 45nm chip micrograph including a 32kb SRAM array and BIST.

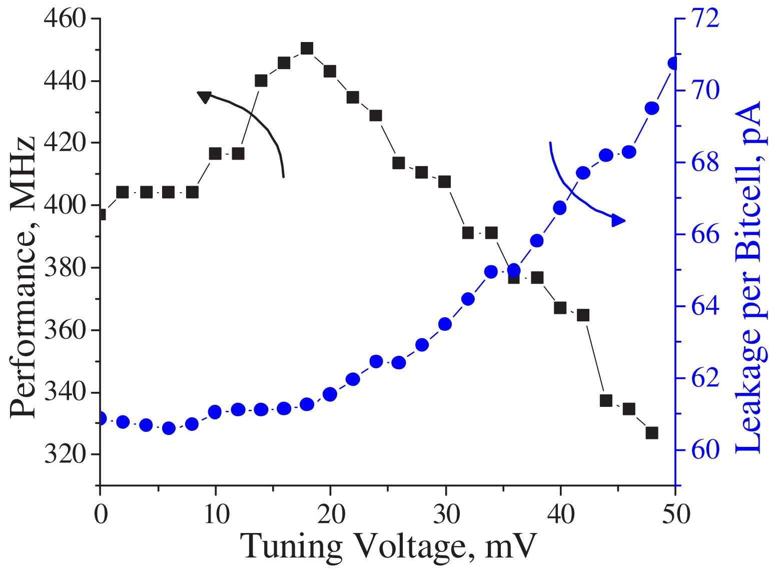


Figure 5.10: Measured results show that Crosshairs can improve array performance by 13% and has a modest leakage overhead.

performance by 13% at a tuning voltage of 20mV (Fig 5.10). It achieves the optimal performance at a tuning voltage that also minimizes the number of simulated and measured stability failures. It creates these performance gains with less than a 2.5% leakage overhead.

5.3.2 Recovering Stability Failures

We record bitcell functionality with no latency requirement for 70 test chips. To measure the impact of Crosshairs on stability, we must first observe some initial PFs. Since nominal PFs are rare, for testing purposes we artificially generate PFs through V_{DD} scaling. Then we recover the resulting PFs using Crosshairs to demonstrate the method's effectiveness. Crosshairs fixes 9 out of 9 initial PFs in an array at a V_{DD} of 532mV with a 26mV tuning voltage (Fig. 5.11). It recovers 15 out of 16 PFs at a V_{DD} of 526mV. The optimal Crosshairs tuning voltage is between 20 and 26mV (Fig. 5.12). These voltages are high enough to fix PFs without creating new failures in tuned rows or columns. In this tuning range, Crosshairs fixes nearly all PFs in arrays with between 1 and 16 initial V_{DD}-scaling-induced PFs. ECC and redundancy can fix a limited number of PFs based on the spatial distribution of failures (Fig. 5.13). Single-error-correct double-error-detect (SECDED) ECC can only fix one PF per word. Higher levels of ECC incur additional area and performance penalties. One measured SRAM array did not yield with ECC. Crosshairs recovers this array

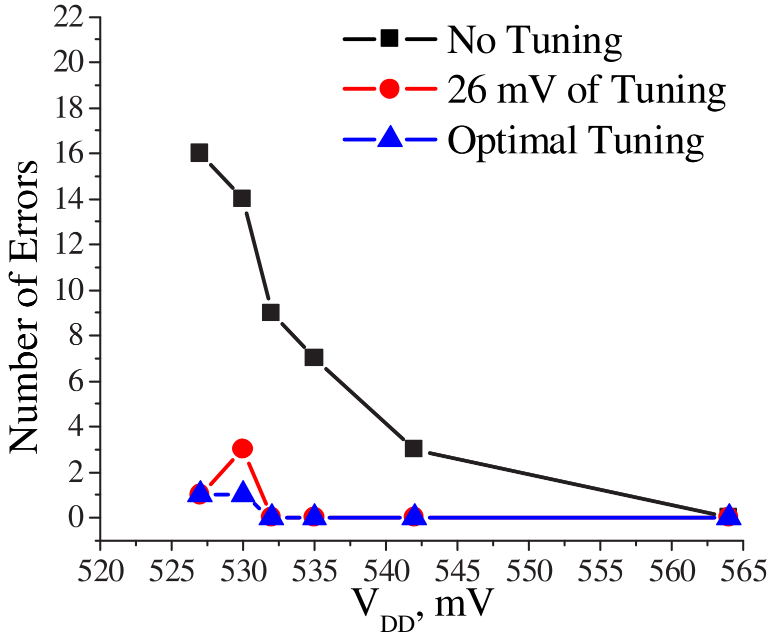


Figure 5.11: Crosshairs recovers measured PFs induced by V_{DD} scaling from the nominal voltage of 1100mV. Using a fixed tuning voltage reduces the overhead for power supply generation with little decrease in effectiveness.

at tuning voltages ranging from 10mV to 50mV. Using redundancy, each column or row with a PF requires an additional redundant row or column, incurring an area and complexity penalty. Crosshairs with a tuning voltage of 26mV reduces the average number of required redundant rows for 100% yield from 4.56 to 1.95. It reduces the required number of redundant columns from 4.46 to 1.91. In addition, the proposed method can be used on top of ECC and redundancy. Fig. 5.14 presents the number of recovered PFs versus initial PFs for all measured chips with a fixed tuning voltage of 26mV. The number of chips at each data point is represented by circle size. Crosshairs fixes an average of 70% of PFs for reasonable initial failure rates lower than 0.1%.

5.4 Conclusions

Crosshairs recovers 70% of PFs in 70 128x256 test arrays by tuning V_{DD} and GND of each SRAM bitcell inverter with respect to its cross-coupled counterpart. These gains are achieved with little modification to a commercial 45nm 6T design and no increase in bitcell area. Crosshairs increases SRAM yield and eliminates or reduces the overheads for other yield improvement techniques such as ECC or redundancy.

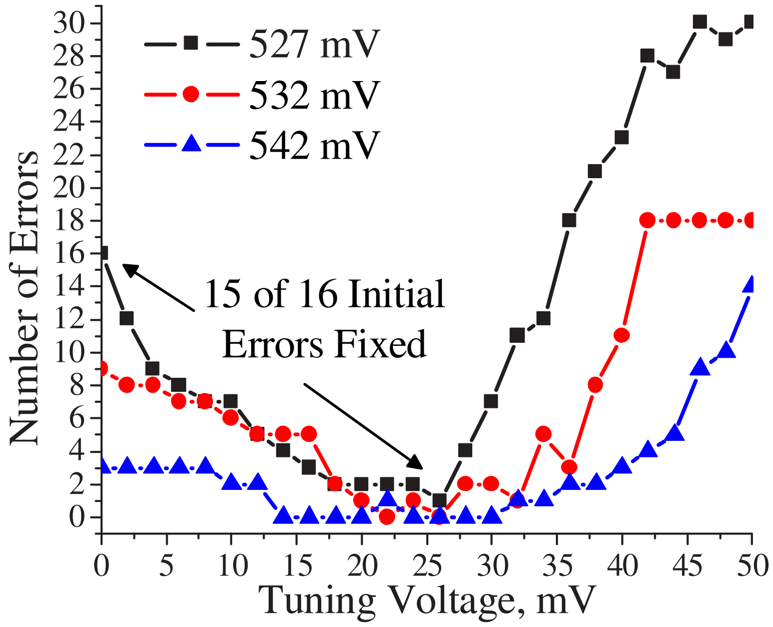


Figure 5.12: The measured optimal tuning voltage is between 20 and 26mV.

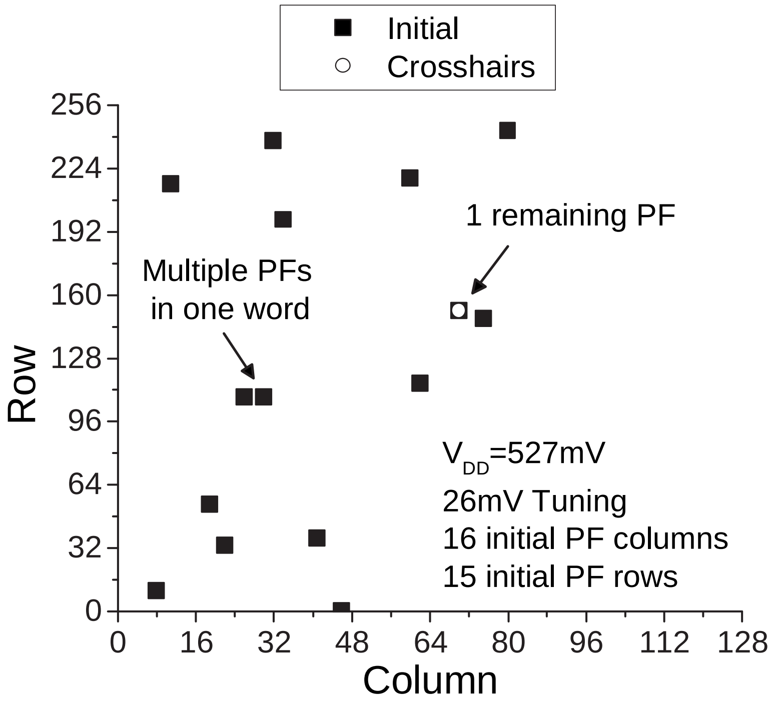


Figure 5.13: Crosshairs can eliminate multiple PFs within one word, unlike SECDED ECC. It can also eliminate PFs spanning many SRAM columns and rows, which would require many redundant columns or rows.

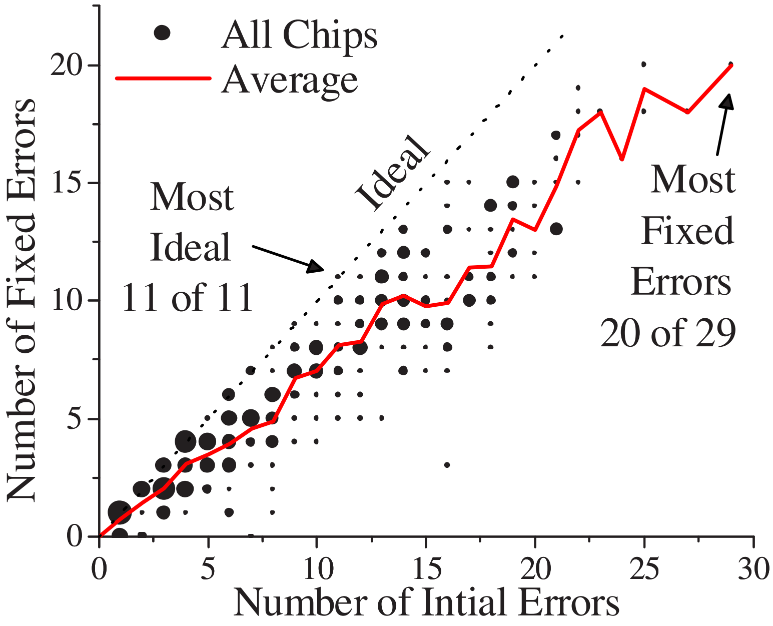


Figure 5.14: Crosshairs reduces the number of PFs in 70 32kb test arrays using a fixed 26mV tuning voltage. It recovers an average of 70% of PFs for reasonable initial failure rates lower than 0.1%.

CHAPTER 6

A Dense 45nm Half-differential SRAM with Improved Voltage Scalability and Operating Margin

6.1 Introduction

In this chapter, we present a robust, low V$_{DD}$ SRAM bitcell for millimeter-scale microsensors with more favorable design tradeoffs than traditional SRAM techniques. Dense, low-power SRAM enables increased data storage and more advanced processing algorithms, while meeting the strict energy budgets of cubic-millimeter systems. Low power techniques such as voltage scaling, which were discussed in Chapter 4, have penalties in robustness, area, timing, or complexity. Many of these penalties arise because of the exacerbated effect of process variations at low V$_{DD}$. Process variations such as random dopant fluctuation and line edge roughness degrade SRAM operating margins, increase parametric failures from timing and stability, and decrease yield [60].

Since designs commonly have large SRAMs, each bitcell must be extremely robust to achieve high chip yield. SRAM designs are typically read-stability limited, so the pull down (PD) must be strong relative to the pass gate (PG) to prevent read upset or destructive read failures. Many process technologies achieve this by using a higher threshold voltage (V$_{TH}$) and longer gate lengths for PGs, in addition to increasing PD width. However, increasing PG V$_{TH}$ and L reduces performance and degrades write margins, creating an upper bound on the limiting stability margin. 8T bitcells separate read and write circuitry to increase margins at the expense of area and leakage [61]. We present HD-SRAM to enable low-power SRAM with lower overheads by achieving high stability margins when the effects of process variations increase. HD-SRAM improves voltage scalability and operating margins with no increase in bitcell size or leakage.

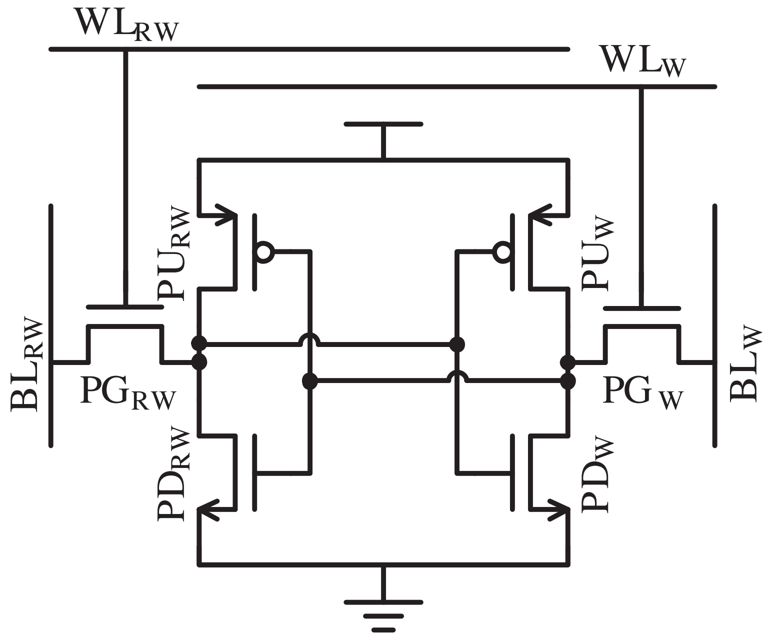
6.2 Half-differential SRAM Method

6.2.1 Operation, Sizing and V$_{TH}$ Selection

HD-SRAM performs a write operation in the same way as D-SRAM. Both PGs turn on and a differential value on the bitlines (BLs) overwrites the cell value. However, during a read operation only one PG turns on and can discharge only one BL (Fig. 6.1). Single-ended read enables asymmetric sizing and V$_{TH}$-selection optimizations to increase robustness without increasing bitcell area. We downsize the write-only pull down device (PDW) because it does not strongly impact read stability. Then, we apply the resulting area savings to increase the read-and-write side PD (PDRW) width and PG (PGRW) length, improving read margin. Since the length of PGRW is increased, we can increase the lengths of PDW and the write-only pull-up device (PUW) no area penalty. This reduces the likelihood of read upsets by decreasing the positive feedback between the cross-coupled inverters and increases the write-one margin. In addition to increasing PDW length, we use the higher NMOS V$_{TH}$ usually reserved for PGs for this device to help prevent read upsets. Using a low V$_{TH}$ device for the write-only pass gate (PGW) further increases write-one margin but decreased the overall simulated robustness and increased leakage, so the device is set to the high V$_{TH}$. Previous asymmetric SRAMs do provide silicon results and either decrease robustness, increase bitcell area, and/or do not consider the physical design of SRAM [62] [63] [64].

6.2.2 Physical Design

The HD-SRAM bitcell has the same area as the commercial D-SRAM bitcell in this 45nm process (0.374μm^2) to allow for an accurate comparison. The layout violates logic design rules to achieve higher density, which is typical for commercial SRAM but uncommon in research efforts [62] (Fig. 6.2). We implemented the design with feedback from the foundry regarding design, lithography, and design for manufacturing (DFM) rules. Two wordlines (WLs) are on Metal 4. All polysilicon is linear and unidirectional to enable double patterning. Unlike most D-SRAM, PDW



Device	W, nm	L, nm	V_{TH}
$\mathbf{PD_{RW}}$	240	55	LVT
$\mathbf{PD_{W}}$	100	95	HVT
$\mathbf{PU_{RW}}$	60	55	SVT
$\mathbf{PU_{W}}$	60	95	SVT
$\mathbf{PG_{RW}}$	150	105	HVT
$\mathbf{PG_{W}}$	100	65	HVT

Figure 6.1: HD-SRAM operates with differential write and single-ended read, enabling asymmetric sizing and V_{TH} selection for higher robustness.

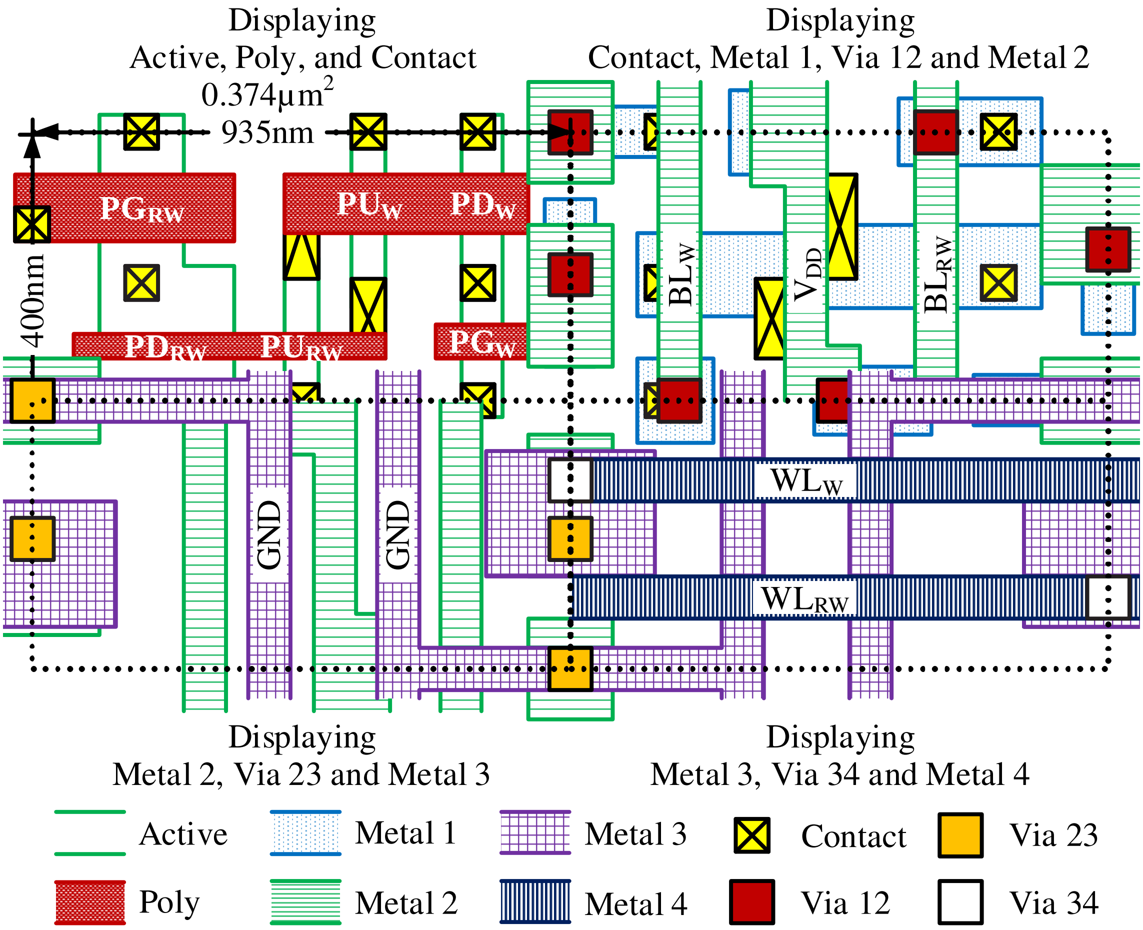


Figure 6.2: HD-SRAM is the same size as a commercial differential 6T design (D-SRAM). Both designs exceed logic design rules for higher density.

and PGW are the same width in HD-SRAM, eliminating a notch in the source-drain region and improving DFM.

6.2.3 Simulated Results

HD-SRAM achieves higher robustness than D-SRAM, even when peripheral assist circuits and optimal technology selection are applied only to D-SRAM. HD-SRAM has a 85-mV higher simulated static noise margin (SNM) than D-SRAM at the nominal V_{DD} of 1.1V (Fig. 6.3a). The HD-SRAM SNM remains higher as V_{DD} scales to below 500mV (Fig. 6.3b).

Since SRAM is typically read-stability limited at nominal V_{DD}, one read assist technique reduces the WL voltage (V_{WL}) to increase read margin [65]. As a measure of robustness, we simulate

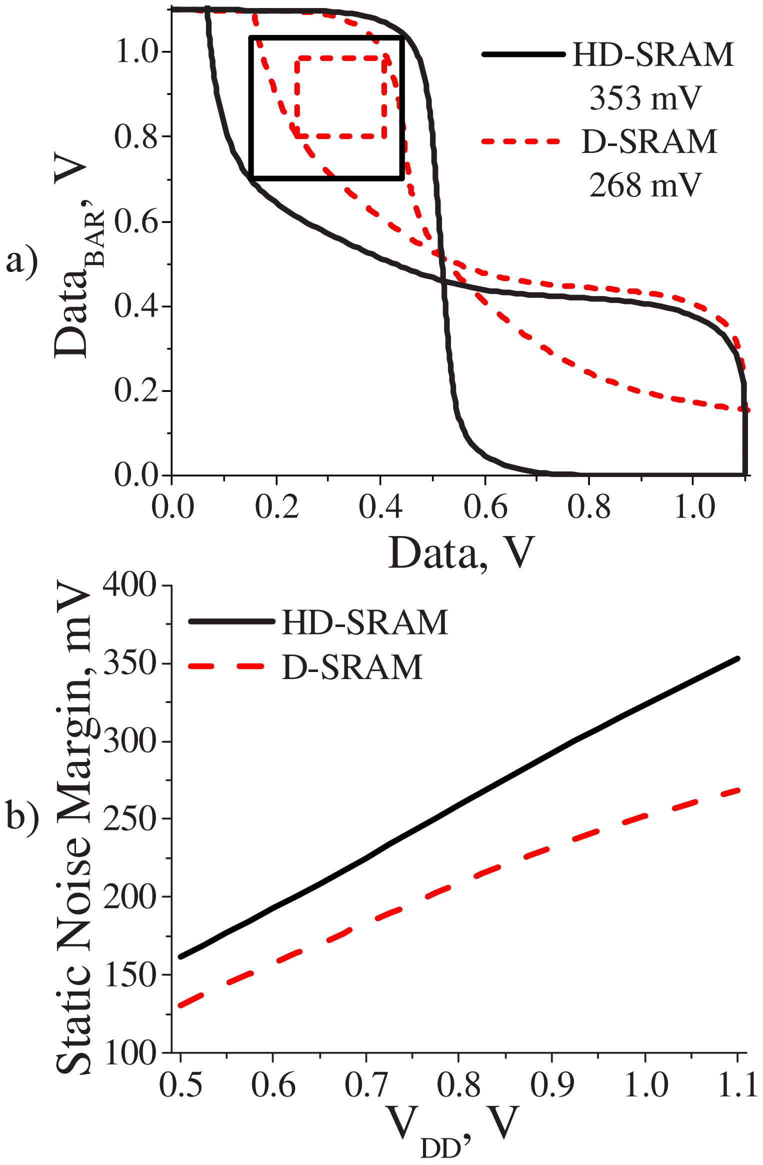


Figure 6.3: HD-SRAM has an 85mV-higher simulated SNM than D-SRAM at nominal V_{DD}. SNM remains higher as V_{DD} scales below 500mV.

the maximum V_{TH} variation that the bitcell can tolerate without functional failure for read, write, and hold operations. We simulate the designs in SPICE using importance sampling and normalize the robustness to a typical 45nm distribution of V_{TH} with σ=40mV [5]. As D-SRAM V_{WL} decreases from 1.1V to 1.02V, read-stability and total robustness increase from 4.2σ to 4.8σ (Fig. 6.4a). However, as V_{WL} further decreases, write margin degrades overall robustness and latency becomes prohibitive. Separate voltages can be used for write and read, but this requires pre-decoding and additional complexity. HD-SRAM without read assistance is more robust than D-SRAM at any V_{WL}. HD-SRAM robustness further improves with read assistance.

The optimal selection of technology parameters, such as V_{TH}, also improves robustness. In typical SRAM processes, these parameters are carefully tuned to optimize the design. However, the nominal V_{TH} selections may trade off robustness for improved performance. We simulate bitcell robustness in SPICE using importance sampling for theoretical selections of technology parameters, with reasonable selections of PD, PG and PU V_{TH}s. The maximum D-SRAM robustness of 4.8σ is achieved by reducing PD and increasing PG V_{TH} (Fig. 6.4b). This robustness is lower than both the nominal and maximum HD-SRAM robustness of 6.1σ of 7.0σ, respectively.

6.3 Measurement Results

6.3.1 Test Chips

We fabricated 32kb banks of HD-SRAM and commercial D-SRAM in a 45nm CMOS process with 1.1V nominal V_{DD} (Figs. 6.5 and 6.6). Each bank uses identical decoders, WL and BL drivers, and sense amplifiers (SAs). HD-SRAM adds gating logic and WL drivers to support the two WLs per row, slightly decreasing array efficiency. We tie one HD-SRAM SA input to a reference voltage to accommodate single-ended read. The test chips do not include assist circuits, error correction coding, or redundancy, which could be applied to either design. A BIST performs functionality and speed tests on each design. A summary of results is presented in Table 6.1.

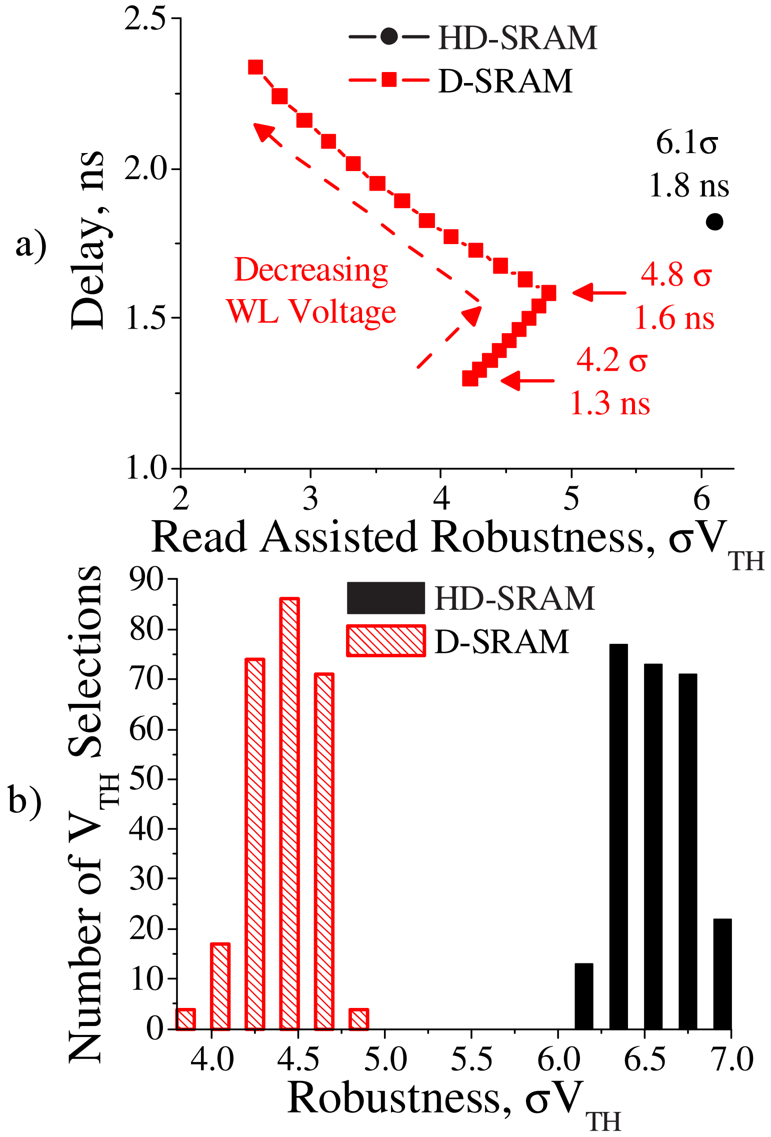


Figure 6.4: D-SRAM robustness improves with assist techniques such as WL voltage selection (a) and technology V_{TH} selection (b). However, neither of these techniques achieves as high robustness as HD-SRAM.

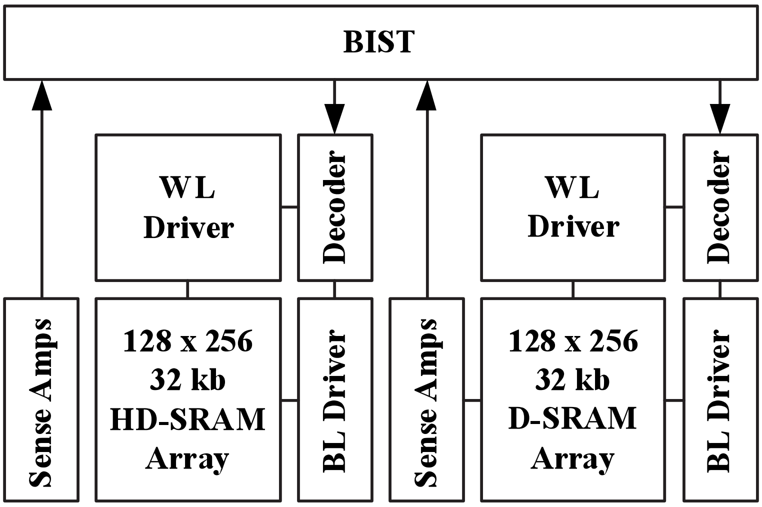


Figure 6.5: HD-SRAM and D-SRAM arrays use nearly-identical peripheral circuits. A BIST performs March and speed tests on both designs.

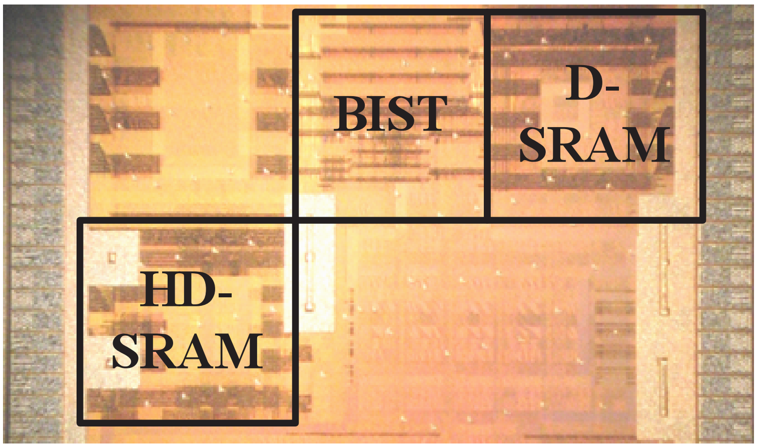


Figure 6.6: HD-SRAM chip micrograph.

Table 6.1: HD-SRAM results summary.

	HD-SRAM	**D-SRAM**
Process	45nm CMOS	45nm CMOS
Bitcell Area	0.37 μm^2	0.37 μm^2
μ **R+W Margin**	12.1 σ	11.0 σ
μ $\mathbf{V}_{MIN}$	639 mV	711 mV
Simulated SNM	353 mV	268 mV
Performance	550 MHz	650 MHz
Energy/bit	43 fJ	53 fJ
Leakage/bit	55 pW	64 pW

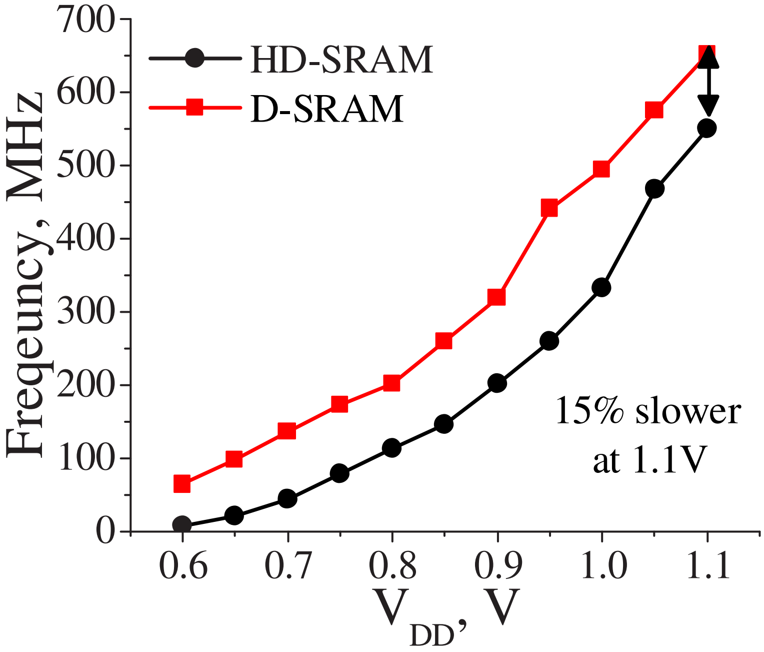


Figure 6.7: Measured performance results show that the HD-SRAM array is 15% slower than D-SRAM at nominal V_{DD}. Array performance is dictated by the slowest cells and HD-SRAM exhibits less timing variation.

6.3.2 Performance, Power and Leakage

D-SRAM is 15% faster than HD-SRAM including WL, bitcell, and BL delays (Fig. 6.7). In a microprocessor, this delay amortizes over register, interconnect, decoder, sense amplifier and multiplexer delays. HD-SRAM has larger read devices that exhibit less timing sensitivity to process variation, decreasing array latency that is dictated by the slowest cells.

HD-SRAM has an 18%-lower access energy than D-SRAM (Fig. 6.8a). HD-SRAM read energy is lower since only one WL switches and capacitance on this WL is lower than the total D-SRAM WL capacitance. Also, in the read-one case, the BL does not discharge. The HD-SRAM write energy is slightly higher because of higher total capacitance on the WLs.

HD-SRAM has a 14%-lower leakage power than D-SRAM (Fig. 6.8b). The leakage improvements result from longer gate lengths selected for PGRW, PDW, and PUW. In addition, PDW has a higher V_{TH} than in D-SRAM.

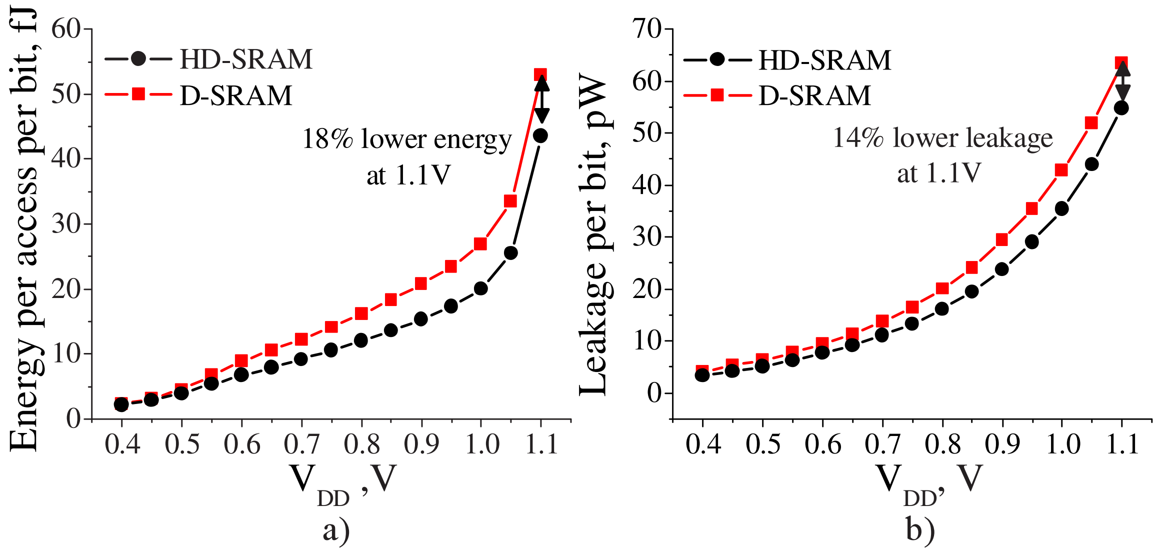


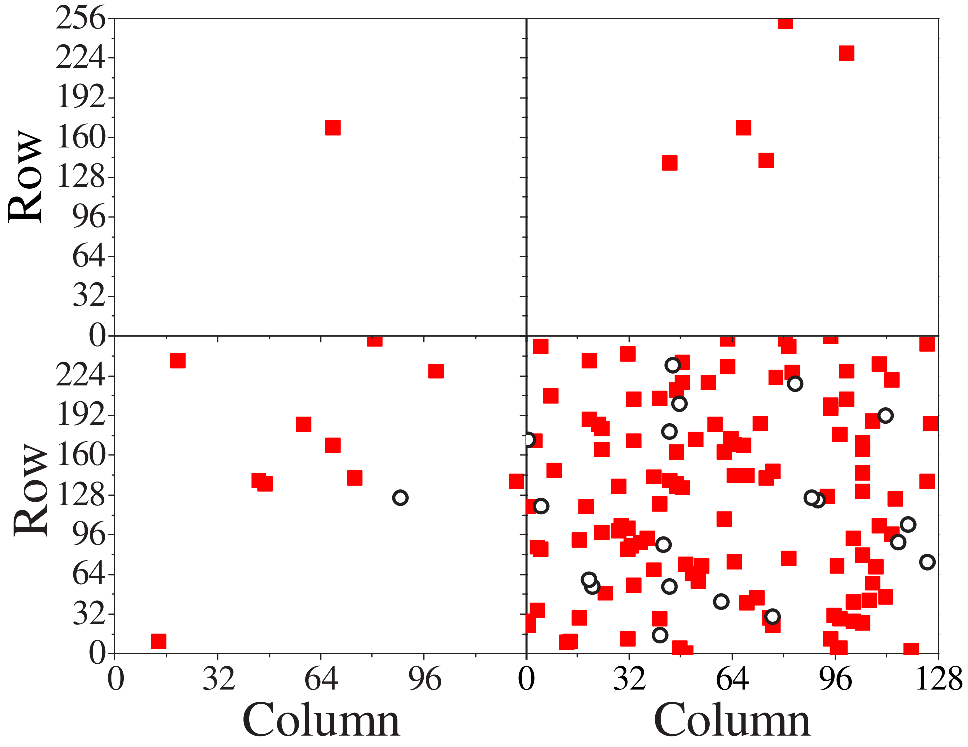
Figure 6.8: HD-SRAM has an 18% lower measured access energy and a 14% lower measured leakage than D-SRAM.

6.3.3 Voltage Scalability

We record the minimum fully-functional V_{DD} (V_{MIN}) and error counts from 80 chips with V_{DD} scaling-induced errors. Error maps from one chip show that V_{MIN} is 800mV for D-SRAM, while every HD-SRAM bitcell functions down to 650mV (Fig. 6.9). Across all 80 test chips, HD-SRAM has an average V_{MIN} that is 72mV lower than D-SRAM (Fig. 6.10). Only 4 HD-SRAM arrays have V_{MIN} above 700mV, whereas 35 D-SRAM arrays fail this criterion. To observe a significant number of errors at V_{DD}s where failures are rare, V_{WL} is raised by 50mV to aggravate read failures. Since these cells are typically read stability limited, this emphasizes variation and emulates cells at the tails of the distribution, which are expected in larger arrays. Under this condition, HD-SRAM has a 100 lower failure rate than D-SRAM (Fig. 6.11).

6.3.4 Operating Margin

To measure read and write margin when few nominal failures are observed, we stress both designs by raising and lowering V_{WL} with respect to V_{DD}, which supplies the cross-coupled inverters (Fig. 6.12). Raising V_{WL} relative to V_{DD} globally aggravates read failures by increasing PG strength and overdrive relative to PD. Similarly lowering V_{WL} reduces PG strength and induces write failures. The shmoo plot in Fig. 6.13 shows that lowering V_{DD} by 25mV causes the first



V_{DD}	○ HD-SRAM Failures	🟥 D-SRAM Failures
800mV	0	0
750mV	0	1
700mV (upper left)	0	1
650mV (upper right)	0	5
600mV (lower left)	1	10
550mV (lower right)	19	115

Figure 6.9: Failure maps show bitcell failure locations as V_{DD} is scaled down. For this test array V_{MIN} is 650mV for HD-SRAM and 800mV for D-SRAM.

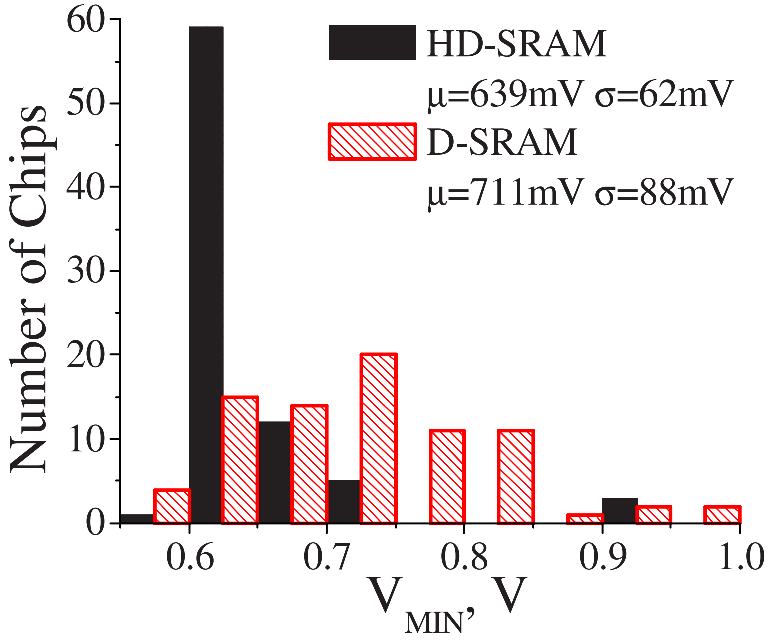


Figure 6.10: A histogram of measured V_{MIN} for 80 test chips shows that HD-SRAM has a 72mV-lower average V_{MIN} and fewer arrays with high V_{MIN}.

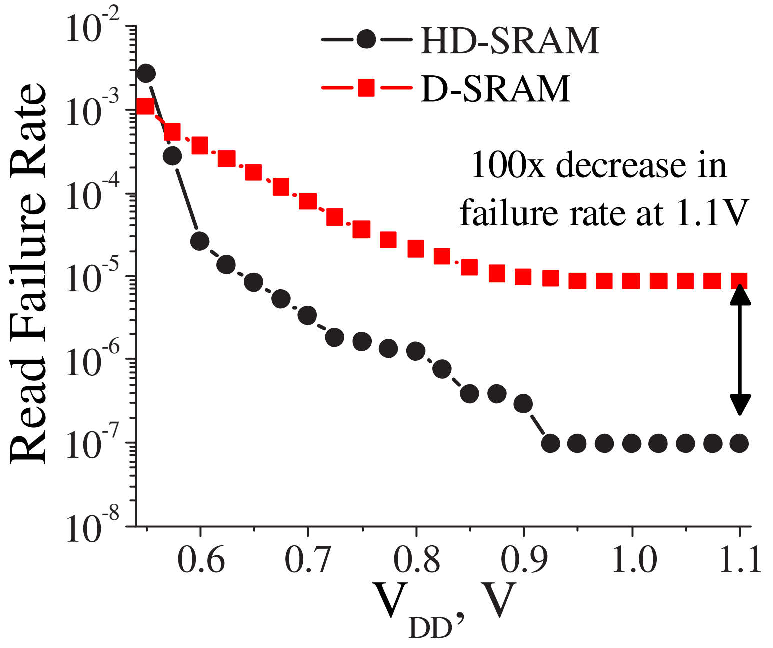


Figure 6.11: At nominal V_{DD}, HD-SRAM has a 100x-lower bitcell failure rate than D-SRAM. Read failures dominate cell stability at nominal V_{DD} and are aggravated in only this plot to observe a significant number of failures.

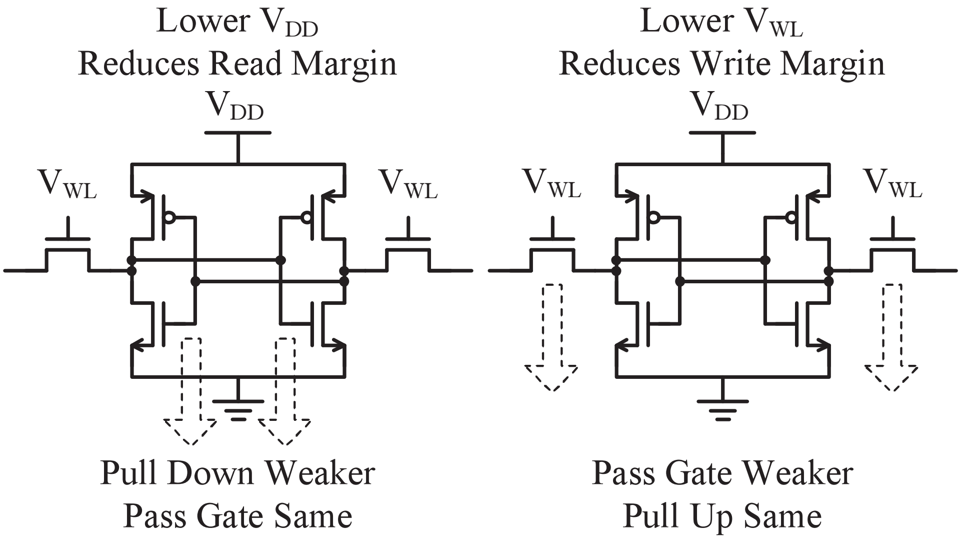


Figure 6.12: Changing the wordline voltage (V_{WL}) with respect to V_{DD} stresses the bitcells to analyze stability when few nominal failures are observed.

D-SRAM array failure when V_{WL} remains fixed at 1.1V. V_{DD} must be decreased by 150mV to cause any HD-SRAM failure. This demonstrates a significantly higher HD-SRAM read margin. Similarly, V_{WL} must be reduced by 375mV and 425mV to cause HD-SRAM and D-SRAM write failures, respectively. The artificial shift in device strengths caused by changing V_{WL} is mapped to an approximate effective-V$_{TH}$ variation tolerance by normalizing the change in device overdrive to the distribution of V$_{TH}$ caused by process variation. This represents the variation a bitcell can tolerate without functional failure. Using a typical 45nm V$_{TH}$ distribution with σV$_{TH}$=40mV, read and write margins are calculated as follows:

$$Margin = \frac{|V_{DD} - V_{WL}|_{1stFailure}}{40mV} \tag{6.1}$$

Over all 80 test chips HD-SRAM exhibits a 2.3σ higher operating margin at 1.1V, which is the minimum of write and read margins and predominantly limited by read stability. However, write margin can be traded for read stability, and vice versa, through device sizing and technology V$_{TH}$ selection, so the sum of read and write margins (R+W) represents a fairer comparison for balanced designs. The average HD-SRAM R+W margin is 1.1σ higher than D-SRAM at 1.1V (Fig. 6.14). HD-SRAM also exhibits a smaller spread in R+W margin, with fewer low margin arrays that fail yield criteria (Fig. 6.15).

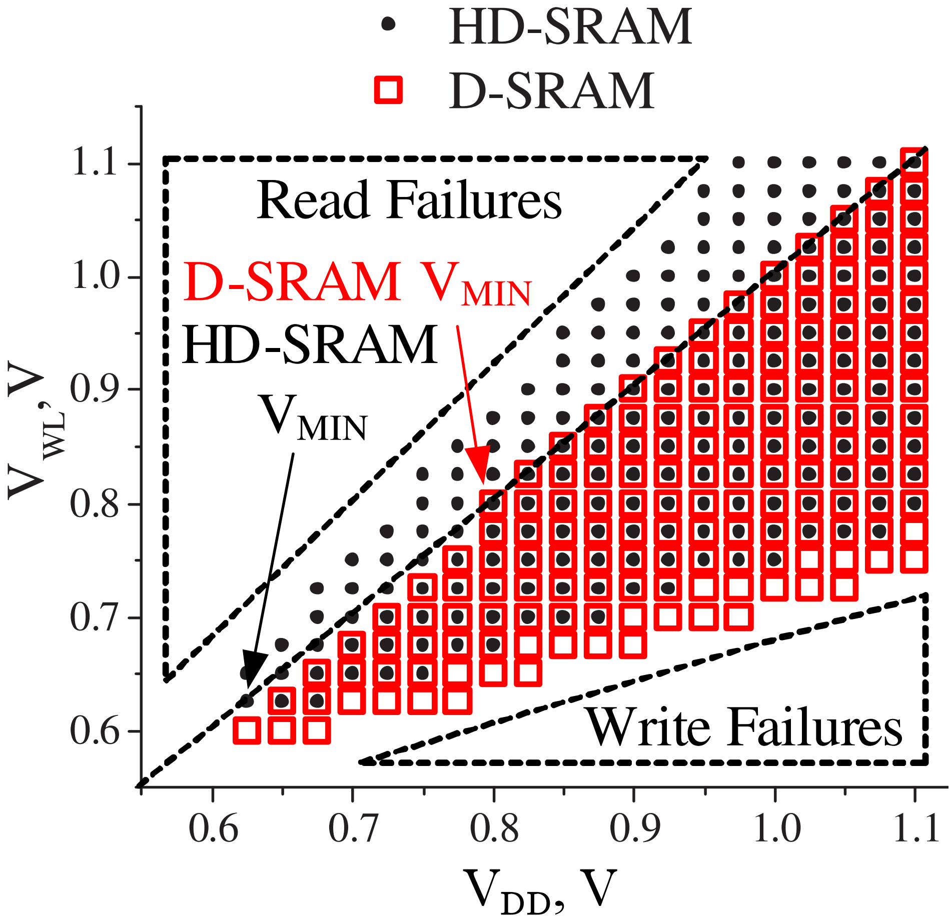


Figure 6.13: Raising and lowering V_{WL} induces artificial mismatch between devices and reduces read and write margins, respectively. The mismatch a cell can tolerate without functional failure is mapped to the R+W margin.

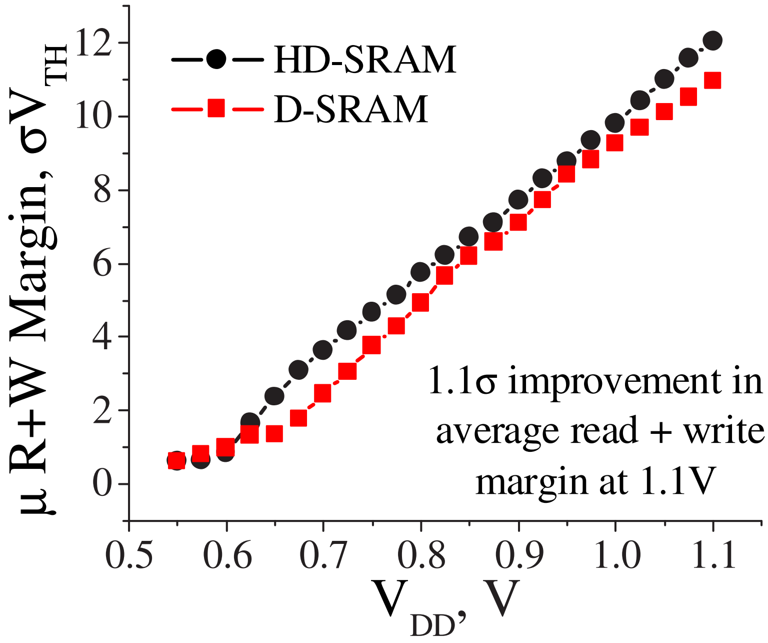


Figure 6.14: HD-SRAM has a 1.1σ-higher R+W margin at nominal V_{DD}.

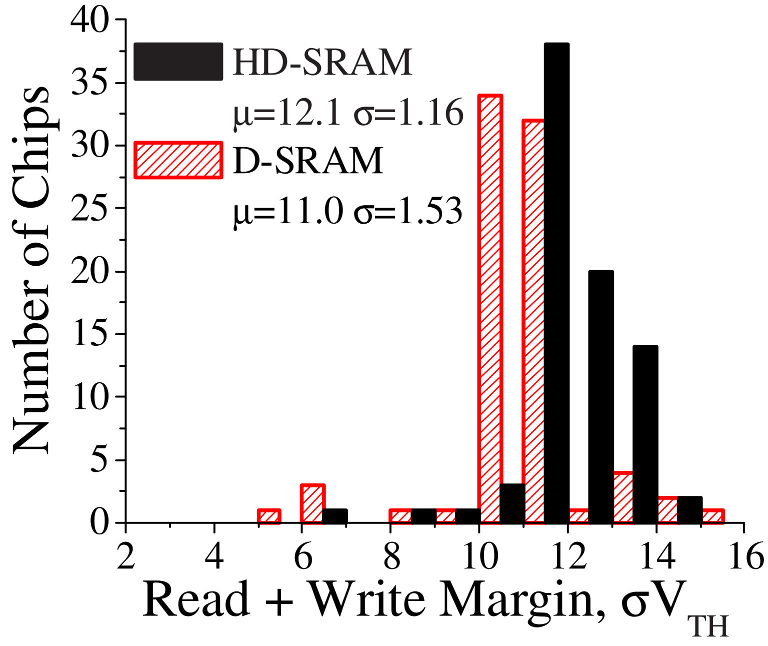


Figure 6.15: HD-SRAM has a tighter distribution of R+W margin, resulting in fewer low-margin arrays that fail yield criteria.

CHAPTER 7

Conclusion

We demonstrate the two first-known, complete, self-powered millimeter-scale computer systems. These microsystems achieve zero-net-energy operation using solar energy harvesting and ultra-low-power circuits. A medical implant for monitoring intraocular pressure (IOP) is presented as part of a treatment for glaucoma. The 1.5mm^3 IOP monitor is easily implantable because of its small size and measures IOP with 0.5mmHg accuracy. It wirelessly transmits data to an external wand while consuming 4.7nJ/bit. This provides rapid feedback about treatment efficacies to decrease physician response time and potentially prevent unnecessary vision loss. A nearly-perpetual temperature sensor is presented that processes data using a 2.1μW near-threshold ARM® Cortex-M3™ μP that provides a widely-used and trusted programming platform.

Energy harvesting and power management techniques for these two microsystems enable energy-autonomous operation. The IOP monitor harvests 80nW of solar power while consuming only 5.3nW, extending lifetime indefinitely. This allows the device to provide medical information for extended periods of time, giving doctors time to converge upon the best glaucoma treatment. The temperature sensor uses on-demand power delivery to improve low-load dc-dc voltage conversion efficiency by 4.75x. It also performs linear regulation to deliver power with low noise, improved load regulation, and tight line regulation.

Low-power high-throughput SRAM techniques help millimeter-scale microsystems meet stringent power budgets. V$_{DD}$ scaling in memory decreases energy per access, but also decreases stability margins. These margins can be improved using sizing, V$_{TH}$ selection, and assist circuits, as well as new bitcell designs. Adaptive Crosshairs modulation of SRAM power supplies fixes

70% of parametric failures. Half-differential SRAM design improves stability, reducing V$_{MIN}$ by 72mV.

The circuit techniques for energy autonomy presented in this dissertation enable millimeter-scale microsystems for medical implants, such as blood pressure and glucose sensors, as well as non-medical applications, such as supply chain and infrastructure monitoring. These pervasive sensors represent the continuation of Bell's Law, which accurately traces the evolution of computers as they have become smaller, more numerous, and more powerful. The development of millimeter-scale massively-deployed ubiquitous computers ensures the continued expansion and profitability of the semiconductor industry. NanoWatt circuit techniques will allow us to meet this next frontier in IC design.

BIBLIOGRAPHY

[1] C.G. Bell, R. Chen, S. Rege, ”Effect of technology on near term computer Structures,” *IEEE Computer*, vol. 5, no. 2, pp. 29-38, Feb. 1972.

[2] Cymbet Corporation, ”Rechargeable thin film battery 12μAh, 3.8V,” EnerChip^{TM} CBC012 datasheet, 2009.

[3] Duracell, ”Alkaline-manganese dioxide battery,” MN1500 datasheet, 2003.

[4] G. Chen, S. Hanson, D. Blaauw, D. Sylvester, ”Circuit design advances for wireless sensing Applications,” *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1808-1827, Nov. 2010.

[5] G.K. Chen, D. Blaauw, T. Mudge, D. Sylvester, and N.S. Kim, “Yield-driven near-threshold SRAM design,” *IEEE/ACM International Conference on Computer-Aided Design*, pp.660-666, 4-8 Nov. 2007.

[6] G. Chen, D. Blaauw, T. Mudge, D. Sylvester, ”Yield-driven near-threshold SRAM design,” *IEEE Transactions on VLSI Systems*, vol. 18, no. 11, pp. 1590-1598, Nov. 2010.

[7] R.G. Dreslinski, G.K. Chen, T. Mudge, D. Blaauw, D. Sylvester, K. Flautner, ”Recongurable energy efficient near threshold cache architectures,” *IEEE/ACM International Symposium on Microarchitecture*, pp.459-470, Nov. 2008.

[8] D. Fick, A. DeOrio, G. Chen, V. Bertacco, D. Sylvester, and D. Blaauw, ”A highly resilient routing algorithm for fault-tolerant NoCs,” *ACM/IEEE Design, Automation, and Test in Europe*, pp. 21-16, Apr. 2009.

[9] M. Wieckowski, G.K. Chen, M. Seok, D. Sylvester, D. Blaauw, ”A hybrid DC-DC converter for nanoampere sub-1V implantable applications,” *IEEE Symposium on VLSI Circuits*, pp. 166-167, June 2009.

[10] G.Chen, M. Fojtik, D. Kim, D. Fick, J. Park, M. Seok, M.-T. Chen, Z. Foo, D. Sylvester, D. Blaauw, ”A millimeter-scale nearly-perpetual sensor system with stacked battery and solar cells,” *IEEE International Solid-State Circuits Conference*, pp. 288-289, Feb. 2010.

[11] M. Seok, S. Hanson, M. Wieckowski, G.K. Chen, Y.-S. Lin, D. Blaauw, D. Sylvester, ”Circuit design advances to enable ubiquitous sensing environments,” *IEEE International Symposium on Circuits and Systems*, pp.285-288, May 2010.

[12] G. Chen, M. Wieckowski, D. Blaauw, D. Sylvester, ”Crosshairs SRAM - an adaptive memory for mitigating parametric failures,” *IEEE European Solid-State Circuits Conference*, pp. 366-369, Sep. 2010.

[13] V. Joshi, M. Wieckowski, G. Chen, D. Blaauw, D. Sylvester, ”Analyzing the impact of double patterning lithography on SRAM variability in 45nm CMOS,” *IEEE Custom Integrated Circuits Conference*, Sep. 2010.

[14] G. Chen, H. Ghaed, R. Haque, M. Wieckowski, Y. Kim, G. Kim, D. Fick, D. Kim, M. Seok, K. Wise, D. Blaauw, D. Sylvester, ”A 1 cubic-millimeter energy-autonomous wireless intraocular pressure monitor,” *IEEE International Solid-State Circuits Conference*, pp.310-311, Feb. 2011.

[15] M. Wieckowski, G. Chen, D. Kim, D. Sylvester, D. Blaauw, ”A 128kb high density portless SRAM using hierarchical bitlines and thyristor sense ampliers,” *IEEE International Symposium on Quality Electronic Design*, in press, 2011.

[16] D. Kim, G. Chen, M. Wieckowski, D. Blaauw, D. Sylvester, ”A 1.85fW/bit ultra low leakage 10T SRAM with speed compensation scheme,” *IEEE International Symposium on Circuits and Systems*, in press, 2011.

[17] G. Chen, M. Wieckowski, D. Kim, D. Blaauw, D. Sylvester, ”A dense 45nm half-dierential SRAM with lower minimum operating voltage,” *IEEE International Symposium on Circuits and Systems*, May 2011.

[18] K.C. Katuri, S. Asrani, and M.K. Ramasubramanian, ”Intraocular pressure monitoring sensors,” *IEEE Sensors Journal*, vol. 8, no. 1, pp. 12-19, Jan. 2008.

[19] P. Wong, C. Kania, M. Shah, M.R. Moinot, J. Schesser, W.C. Hunter, T.L. Alvarez, ”Interfacing a tonometer with a microcontroller to monitor diurnal intraocular pressure variations,” *IEEE Northeast Bioengineering Conference*, Mar. 2010.

[20] M. Leonardi, P. Leuenberger, D. Bertrand, A. Bertsch, P. Renaud, ”A soft contact lens with a MEMS strain gage embedded for intraocular pressure monitoring,” *IEEE International Conference on Transducers, Solid-State Sensors, Actuators, and Microsystems*, vol. 2, pp. 1043-1046, June 2003.

[21] K. Stangel, S. Kolnsberg, D. Hammerschmidt, B.J. Hosticka, H.K. Trieu, W. Mokwa, ”A programmable intraocular CMOS pressure sensor system implant,” *IEEE Journal of Solid State Circuits*, vol. 36, no. 7, pp. 1094-1100, July 2001.

[22] P.-J. Chen, D.C. Rodger, S. Saati, M.S. Humayun, Y.-C. Tai, ”Microfabricated implantable parylene-based wireless passive intraocular pressure sensors,” *IEEE Journal of Microelectromechanical Systems*, vol. 17, no. 6, pp. 1342-1351, Dec. 2008.

[23] Y.-C. Shih, T. Shen; B. Otis, ”A 2.3μW wireless intraocular pressure/temperature monitor,” *IEEE Asian Solid State Circuits Conference*, Nov. 2010.

[24] E. Y. Chow, S. Chakraborty, W.J. Chappell, P.P. Irazoqui, ”Mixed-signal integrated circuits for self-contained sub-cubic millimeter biomedical implants,” *IEEE Internation Solid-State Circuits Conference*, pp. 236-237, Feb. 2010.

[25] R.M. Haque, K.D. Wise, ”A 3D implantable microsystem for intraocular pressure monitoring using a glass-in-silicon reflow process,” *IEEE International Conference on Micro Electro Mechanical Systems*, pp. 995-998, Jan. 2011.

[26] D.G. Vaughn, T. Asbury, P. Riordan-Eva, *General Ophthalmology*, 15th ed. Stamford, CT: Appleton & Lange, 1999.

[27] V. Raghunathan, A. Kansal, J. Hsu, J. Friedman, M. Srivastava, ”Design considerations for solar energy harvesting wireless embedded systems,” *International Symposium on Information Processing in Sensor Networks*, pp. 457-462, April 2005.

[28] E. Carlson, K. Strunz, B. Otis, ”20mV input boost converter for thermoelectric energy harvesting,” *IEEE Symposium on VLSI Circuits*, pp. 162-163, June 2009.

[29] M.D. Seeman, S.R. Sanders, J.M. Rabaey, ”An ultra-low-power power management IC for energy-scavenged Wireless Sensor Nodes,” *IEEE Power Electronics Specialists Conference*, pp. 925-931, June 2008.

[30] Y.K. Ramadass, A.P. Chandrakasan, ”Minimum energy tracking loop with embedded DC-DC converter enabling ultra-low-voltage opperation down to 250 mV in 65 nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 256-265, Jan. 2008.

[31] J. Kwong, Y.K. Ramadass, N. Verma, A.P. Chandrakasan, ”A 65 nm sub-Vt microcontroller with integrated SRAM and switched capacitor DC-DC converter,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 115-126, Jan. 2009.

[32] M.D. Seeman, S.R. Sanders, ”Analysis and optimization of switched-capacitor DC-DC converters,” *IEEE Transactions on Power Electronics*, vol. 23, no. 2, pp.841-451, Mar. 2008.

[33] G.K. Balachandran, R.E. Barnett, ”A 110 nA voltage regulator system With dynamic bandwidth boosting for RFID systems,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 9, pp. 2019-2028, Sept. 2006.

[34] Y.K. Ramadass, A.P. Chandrakasan, ”Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications,” *Power Electronics Specialists Conference*, pp. 2353-2359, June 2007.

[35] M. Seok, G. Kim, D. Sylvester, D. Blaauw, ”A 0.5V 2.2pW 2-transistor voltage reference,” *IEEE Custom Integrated Circuits Conference*, pp. 577-580, Sept. 2009.

[36] C. Riordan, R. Hulstron, ”What is an air mass 1.5 spectrum? [solar cell performance calculations],” *IEEE Photovoltaic Specialists Conference*, vol. 2, pp. 21-25, May 1990.

[37] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, “Theoretical and practical limits of dynamic voltage scaling.”, c*ACM/IEEE Design Automation Conference*, pp. 868-873, May 2004.

[38] A. Wang and A. Chandrakasan, “A 180mV FFT processor using subthreshold circuit techniques,” *IEEE International Solid-State Circuits Conference*, vol. 1, pp. 292-529, Feb. 2004.

[39] N. Lindert, T. Sugii, S. Tang, and C. Hu, “Dynamic threshold pass-transistor logic for improved delay at lower power supply voltages,” *IEEE Journal of Solid-State Circuits*, vol. 34, num. 1, pp. 85-89, Jan. 1999.

[40] E. Seevinck, F.J. List, and J. Lohstroh, “Static-noise margin analysis of MOS SRAM cells,” *IEEE Journal of Solid-State Circuits*, vol. 22, num. 5, pp. 748-754, Oct. 1987.

[41] M.J.M. Pelgrom, H.P. Tuinhout, and M. Vertregt, “Transistor matching in analog CMOS applications,” *IEEE International Electron Devices Meeting*, pp. 915-918, Dec. 1998.

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

[43] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, and H. Shinohara, "A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits," *IEEE Journal of Solid-State Circuits*, vol. 42, num. 4, pp. 820-829, Apr. 2007.

[44] O. Hirabayashi, A. Kawasumi, A. Suzuki, Y. Takeyama, K. Kushida, T. Sasaki, A. Katayama, G. Fukano, Y. Fujimura, T. Nakazato, Y. Shizuki, N. Kushiyama and T. Yabe, "A process-variation-tolerant dual-power-supply SRAM with $0.179\mu m^2$ cell in 40nm CMOS using level-programmable wordline driver," *IEEE International Solid-State Circuits Conference*, pp. 458-459, Feb. 2009.

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

[46] H. Pilo, C. Barwin, G. Braceras, C. Browning, S. Lamphier, and F. Towler, "An SRAM design in 65-nm technology node featuring read and write-assist circuits to expand operating voltage," *IEEE Journal of Solid-State Circuits*, vol. 42, num. 4, pp. 813-819, Apr. 2007.

[47] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200mV 6T SRAM in 0.13μm CMOS," *IEEE International Solid-State Circuits Conference*, pp. 332-606, Feb. 2007.

[48] B.H. Calhoun and A. Chandrakasan, "A 256kb sub-threshold SRAM in 65nm CMOS," *IEEE International Solid-State Circuits Conference*, pp. 628-629, Feb. 2006.

[49] T.-H. Kim, J. Liu, J. Keane, and C.H. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," *IEEE International Solid-State Circuits Conference*, pp. 330-606, Feb. 2007.

[50] J.P. Kulkarni, K. Kim, and K. Roy, "A 160mV robust Schmitt Trigger based subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 42, num. 10, pp. 2303-2313, Oct. 2007.

[51] C.H. Kim and K. Roy, "Dynamic Vt SRAM: a leakage tolerant cache memory for low voltage microprocessors," *IEEE International Symposium on Low Power Electronics and Design*, pp. 251-254, Aug. 2002.

[52] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, num. 12, pp. 1859-1880, Dec. 2005.

[53] P. Shahabuddin, "Importance sampling for the simulation of highly reliable Markovian systems," *Management Science*, vol. 40, num. 3, pp. 333-352, Mar. 1994.

[54] R. Kanj, R. Joshi, and S. Nassif, “Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events,” *ACM/IEEE Design Automation Conference*, pp. 69-72, Jul. 2006.

[55] P. Heidelberger, “Fast simulation of rare events in queueing and reliability models,” *ACM Transactions on Modeling and Computer Simulation*, vol. 5, num. 1, pp. 43-85, Jan. 1995.

[56] A. Agarwal, B.C. Paul, S. Mukhopadhyay, and K. Roy, “Process variation in embedded memories: failure analysis and variation aware architecture,” *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp.1804-1814, Sept. 2005.

[57] K. Jeong, and A.B. Kahng, “Timing analysis and optimization implications of bimodal CD distribution in double patterning lithography,” *IEEE Asia and South Pacific Design Automation Conference*, pp.486-491, Jan. 2009.

[58] C.L. Chen, and M.Y. Hsiao, “Error-correcting codes for semiconductor memory applications: A state-of-the-art review,” *IBM Journal of Research and Development*, vol.28, no.2, pp.124-134, Mar. 1984.

[59] J. P. Bickford, R. Rosner, E. Hedberg, J.W. Yoder, and T.S. Barnett, “SRAM redundancy - Silicon area versus number of repairs tradeoff,” *Advanced Semiconductor Manufacturing Conference*, pp.387-392, May 2008.

[60] R. Aitken, and S. Idgunji, “Worst-case design and margin for embedded SRAM,” *Design, Automation and Test in Europe Conference*, pp.1-6, Apr. 2007.

[61] L. Chang, R.K. Montoye, Y. Nakamura, K.A. Batson, R.J. Eickemeyer, R.H. Dennard, W. Haensch, and D. Jamsek, “An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches,” *IEEE Journal of Solid-State Circuits*. vol. 43, no. 4, pp.956-963, Apr. 2008.

[62] N. Azizi, F.N. Najm, andA. Moshovos, “Low-leakage asymmetric-cell SRAM,” *IEEE Transactions on VLSI*, vol.11, no.4, pp.700-715, Aug. 2003.

[63] B.S. Gill, C. Papachristou, and F.G. Wolff, “A new asymmetric SRAM cell to reduce soft errors and leakage power in FPGA,” *Design, Automation and Test in Europe Conference*, pp.1-6, Apr. 2007.

[64] K. Kim, J.-J. Kim, C.-T. Chuang, “Asymmetrical SRAM cells with enhanced read and write margins,”, *IEEE International Symposium on VLSI Technology, Systems and Applications*, pp.1-2, Apr. 2007.

[65] K. Nii, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, H. Makino, Y. Yamagami, S. Ishikura, T. Terano, T. Oashi, K. Hashimoto, A. Sebe, S. Okazaki, K. Satomi, H. Akamatsu, H. Shinohara, “A 45-nm bulk CMOS embedded SRAM With improved immunity against process and temperature variations,” *IEEE Journal of Solid-State Circuits*. vol. 43, vol. 1, pp.956-963, Jan. 2008.