Improving Reliability and Power Consumption of Memories in Battery-less Systems-on-Chip

---

A Dissertation

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Doctor of Philosophy (Electrical Engineering)

by

Farah B. Yahya

January 2017

Abstract

Recent projections by Cisco have suggested that more than 50 billion devices are expected to sense, process and transmit information as part of the internet of things (IoT) by 2020. These devices will target a wide range of applications including but not limited to health monitoring, environmental monitoring, infrastructure monitoring, smart homes, and smart cars. Due to the varying environmental conditions under which these devices are expected to operate and to their large number, battery replacement has become a major concern. Recently, ultra-low power (ULP) systems-on-chip (SoCs) that can operate solely on harvested energy have been presented. A number of challenges face the dispersion of this technology: 1) reducing the power and energy consumption of the building blocks of these SoCs to stay within the budget of the energy harvester, 2) operating reliably under varying harvesting conditions, and 3) maintaining critical data in the event of a power loss. This work will investigate techniques to address these challenges and enable battery-less operation.

One of the main techniques used to reduce power consumption is scaling the supply voltage (V_{DD}) below the threshold voltage of the transistors (sub-threshold operation). Due to the quadratic dependence of power on V_{DD}, Reducing V_{DD} to the sub-threshold domain will result in significant power savings thus enabling battery-less operation. However, with supply scaling comes the additional challenges of reduced reliability and performance. While performance is an acceptable trade-off in applications with low throughput requirements, ensuring reliability at low voltages remains a challenge. As V_{DD} is scaled to the sub-threshold region, the on-to-off current ratio of a transistor is reduced, and the impact of process

variations on its strength increases. These trends cause increased failures in ratio-ed circuits - such as the widely used static random access memory (SRAM) bit-cell - that depend on the relative strengths of their transistors for correct operation.

SRAM cells are volatile in nature and thus lose their data when power is lost. However, they consume lower read and write energy than their non-volatile counterparts do. Emerging non-volatile cells such as spin-torque transfer RAM (STT-RAM) and ferroelectric RAM (FeRAM) have been introduced as replacements for the high power FLASH memories. While these cells consume significantly lower power than FLASH, their power consumption is still considerably higher than SRAMs which limits their use in battery-less devices.

In this work, we will present an ULP battery-less IoT system-on-chip (SoC) with a non-volatile auto-recovery backup sub-system and highly optimized components to allow operation within a sub-μW power budget. Techniques are introduced within the SRAM instruction and data memories to improve their reliability and reduce their power consumption. A backup sub-system containing FeRAM arrays is also implemented to help the SoC recover in the event of power loss. A ULP voice activity detector is also developed as an always-on wake-up for the SoC. The combination of these techniques enable reliable battery-less operation of the SoC.

Approval Sheet

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Electrical Engineering)

Farah B. Yahya

This dissertation has been read and approved by the Examining Committee:

Benton Calhoun, Adviser

Joanne Dugan, Committee Chair

John Lach

Kamin Whitehouse

Steven Bowers

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

January 2017

To everyone who's helped me succeed

Acknowledgements

Throughout my four years in graduate school, I had the honor and privilege of working alongside a group of extremely intelligent and innovative people. First and foremost, Professor Benton Calhoun - my advisor - is a continuous source of inspiration. His positive attitude and his enthusiasm towards building circuits is un-rivaled. He sees the best in all his students and drives us to always seek excellence in our work and life. Thanks for being a great mentor!

Special thanks to my committee: Professors Joanne Dugan, John Lach, Kamin Whitehouse, and Steven Bowers. Their constructive criticism of my work has helped improve its quality significantly. Professor Dugan has always been a model of what female engineers should aspire to be. Professor Lach provided constant feedback on building self-powered systems and helped direct many of the decisions during the design process. Professor Bowers has been very patient with my constant questions about analog circuits. Professor Whitehouse always provided big picture comments that help direct the motivation of the work and find useful applications that could benefit from different aspects of the work. Thank you for being very supportive and positive.

Bengroup alumni have made grad school wonderful for me. Alicia Klinefelter is a close friend and a model graduate student: very knowledgeable, very smart, and very supportive. Yousef Shakhsheer is a constant support, ready to help with technical and non-technical matters. Without Kyle Craig's awesome scripts, grad school would have been significantly more challenging. Jim Boley's SRAM and TASE tests made my first year in grad school run a lot smoother. Oluseyi Ayorinde's enthusiasm and positive attitude were a breath of fresh

air during deadlines. Thank you all for being awesome friends!

I came in to grad school with three super-smart and super-friendly students: Harsh Patel, Abhishek Roy, and Christopher Lukas. We have collaborated on many of the projects presented in this thesis and a lot others. Thank you for being model collaborators. Thank you for being patient with me. Thank you for all the lunches, coffee breaks, discussions, and late night tapeout help. You have all been awesome to work with and I have enjoyed every moment of it.

I also would like to thank current Bengroup members: He Qi, Divya Akella, Arijit Banarjee, Ningxi Lui, Shuo Li, Daniel Trusdell, and Kevin Leach. A special thanks to Jacob Breiholz for always being available to test the different chips we worked on. His methodical testing has made it possible for me to focus more on tapeouts while still making sure that the designs are behaving as expected. Lach's group - Luis Lopez Ruiz, Jackey Gong, Matthew Ridder, and Ben Ghaemmaghami - has also been super helpful with ASSIST demos. I have also had the privilege of working with Professor Dave Wentzloff and his students - Xing Chen and Avish Kosari - at the University of Michigan to develop complex systems. Xing has been especially wonderful when we visited Michigan: showing us around and making sure we had everything we needed.

During the last year, I had the opportunity to work with an awesome group of people at Texas Instruments: Steven Bartling, Sudhanshu Khanna, Wendy Barr, Hemalata Sangodkar, Scott Summerfelt, and Zakir Shaik. Thank you for being patient during the PDK setup and for taking the time out of your busy schedules to support us. My managers, Muhammad Khellah and Vikas Chandra, have been great mentors during my internships at Intel and ARM. Their technical advice has helped me grow as a professional. I also would like to thank the funding agencies and UVA for enabling all the work done in this dissertation.

I do not think there are enough words to describe my gratitude and love to Terry Tigner and her wonderful family. They have welcomed me into their home and life and treated me like a daughter. They are my family in the US.

My friends from Lebanon: Doa'a Al Otoom, Nizar Zarif, and Jad Ghandour have been my go-to people when good news came along and my shoulder to cry on when things got rough. They always encouraged me and listened to my nagging. They provided their comforting advice without reserve. Thank you for being the amazing people you are!

I could not have made it this far in life without the unconditional love of my family. They have always cheered me on and encouraged me to be the best I could be, even when that path took me away from home. Without my mom's wonderful recipes and her wide knowledge in gardening and interior design, my life would have been much less colorful. Basima's wise advice on how to handle conflict and stay calm has been essential in my success. If you find a chapter in my dissertation or a section in my paper without errors and with flawless English, you'll directly know that Haya went over it. Fadi's advice on life in the US taught me a lot and made my transition so much easier. Finally, I owe all my successes in life to my personal hero: my dad. He has been a constant example of what success, modesty, kindness, intelligence, and integrity are like. My love of learning is something I have inherited from him. No words can express my love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

More and more devices are being deployed as part of the internet-of-things (IoT) in a continuous effort to improve the quality of human life. These devices range from wearables that monitor physiological signal (such as electrocardiograph - ECG) - in order to detect critical conditions - to devices designed to sense environmental conditions and infrastructure integrity - to provide early warnings of potential hazards. Other devices, designed to enable next generation autonomous cars, also sense and gather data in an attempt to reduce the number of road accidents. Smart sensors around the house improve safety and security, and provide more control and monitoring.

With such broad range of applications expected from IoT devices, many challenges and limitations are facing IoT device designers. In applications such as body sensors, form factor and battery lifetime are major limitations. Environmental and Infrastructure sensors are expected to be power autonomous since they will be spread out geographically. While batteries can enable power autonomy, they require continuous recharging and have limited number of recharge cycles forcing regular replacement. Thus, devices that can power themselves by harvesting energy from their environments become particularly attractive in the context of

Table 1.1: Energy sources and their corresponding harvesting power [1]

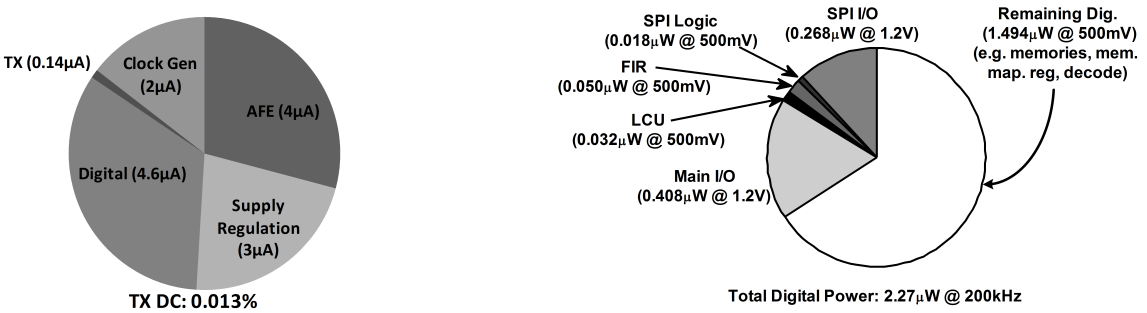
Energy Source		Harvested Power $/cm^2$
Light	Indoor	10μW
	Outdoor	10mW
Thermoelectric	Human	4μW
	Industrial	100μW
Vibration	Human	30μW
	Industrial	1-10mW
RF		0.1μW

IoT. Energy harvesting battery-less IoT devices provide a number of advantages: they reduce the burden of battery replacement, they are more environmentally friendly, and they can be used to improve the quality of life in developing countries where a continuous power source is not available to everybody.

Currently, battery powered solutions are still dominant. However, a number of battery-less systems-on-chip (SoCs) [2][3][4] have recently been introduced targeting IoT applications. These SoCs have employed different energy sources as their main power supply. Table 1.1 summarizes the most widely used sources and the amount of power that can be harvested from each given a 1 cm^2 harvester. The table shows that the amount of harvested power is limited and varies significantly with the surrounding environment. This amount also depends on other factors including time of day/year and ambient temperature. Thus, due to the limited and varying power available for battery-less SoCs, their building blocks must be carefully designed to reduce power consumption while maintaining reliability. They must also adapt to the varying harvesting conditions, and recover from complete power loss.

This dissertation addresses the three main challenges facing battery-less operation, namely: power, reliability and recovery. To reduce power, the main contributors to the power budget are identified. Figure 1.1 shows a breakdown of the power consumption of two recent battery-less SoCs. The digital processing component including the on-chip memories account for the majority of the power budget. Sensing interfaces - such as the analog front end (AFE) - also consumes a significant portion of the power budget. Scaling the supply voltage of these

components is one of the main techniques used to reduce their power consumption. However, aggressive supply scaling raises reliability concerns especially in components that depend on the relative strength of their devices for reliable operation. One such circuit is the on-chip static random access memory (SRAM). Due to their large number, SRAMs also suffer from variations due to manufacturing which impacts their reliability. Non-volatile memories - especially emerging ones like spin-transfer torque (STT) and ferroelectric (Fe) RAM - also suffer from manufacturing variation. These memories are essential to enable recovery from power loss, since - unlike SRAMs - they are capable of retaining their data in the event of complete power loss.



(a) Current consumption reported in [2] (b) Power consumption reported in [3]

Figure 1.1: Power distribution in recent chips.

1.2 Thesis

To enable the trillion node IoT vision, self-powered SoCs must operate reliably on a limited and continuously varying power budget. Voltage scaling to the sub-threshold region reduces power consumption at the expense of reliability. Thus, novel circuits techniques must be introduced into each building block of the SoC to improve its reliability. At the same time, these blocks must be co-designed to ensure low power operation. Continuous power monitoring combined with a non-volatile memory will allow the SoC to adapt to the varying power budget and recover from power failures.

1.2.1 Reducing Power

A suite of power saving techniques must be introduced into the major consumers on these systems. These blocks must make use of circuits techniques such as scaling and completely gating the supply voltage, using low power devices and duty cycling to minimize their power consumption. In addition, architectural optimization and component co-design will enable further reduction in power.

1.2.2 Improving Reliability

Introducing different assist techniques into the on-chip memories will help maintain their reliability especially at low supply voltages. Eliminating reference generators in on-chip and off-chip memories will also reduce the impact of manufacturing variations.

1.2.3 Enabling Recovery

A non-volatile memory sub-system designed specifically to interface to battery-less SoCs will enable recovery from complete power loss when harvesting conditions are non-ideal. However, different circuits and architectural techniques are needed to ensure reliable and low power operation of this recovery sub-system.

1.3 Approach

To address the three main concerns, we look into three main components of battery-less SoCs: the on-chip SRAM memories, the off-chip non-volatile memories and an example always-on sensing interface.

1.3.1 Reducing the SRAM Power Consumption

SRAMs hold instructions and data within a battery-less SoC and thus are indispensable. However, they suffer from reliability concerns especially when operated at low voltages. Thus, reducing the supply voltage of SRAM to reduce their power consumption must be studied carefully to avoid compromising reliability. In this approach, we study the impact of voltage scaling on reliable read and write operations. Different read and write assist techniques are also evaluated to determine the optimal combination that will enable reliable low voltage operation. Two bit-cells are considered for low voltage low power operation. Based on this study, an SRAM array is built and fabricated in a 130nm CMOS process. This array can operate reliability at low voltages (down to 350mV), and takes advantages of many power saving techniques to enable retention at a minimum of 12.29nW/KB. The main controller on the SoC is then modified to take advantage of the different low power features of the SRAM to significantly cut down the power consumption of the system.

1.3.2 Introducing a Non-Volatile Back-up Sub-system

In this approach, we study two potential emerging non-volatile memories for their use in battery-less SoCs. Both STT-RAM and Fe-RAM can retain their data without consuming any power, however, their read and write energy is considerably larger than SRAMs. Thus, we look into different techniques to reduce their read/write energy. Then, we develop an Fe-RAM based back-up sub-system to complement a battery-less SoC. The SoC will request data from the Fe-RAM sub-system upon a power-on reset. This data is then read by a cold-boot circuit on the SoC and used to program the on-chip memory. The back-up sub-system also consists of a first-in first-out (FIFO) array that the SoC can use to save critical system data.

1.3.3 Designing an Ultra-Low-Power Sensing Interface

Since the digital processing and sensing interfaces consume a significant portion of the power consumed by battery-less SoCs, in this approach, we choose to design a ULP always-on voice activity detector. We look into computationally inexpensive algorithms to detect voice activity and then study different techniques to improve their accuracy and reduce their power consumption. A simple analog front end is developed to assist the digital algorithm to achieve a ULP wake-up sensing interface. This wake-up interface is co-designed with the main controller of the SoC to allow the entire system to remain in an extremely low power (<50nW) state until a wake-up system is detected.

1.4 Dissertation Contributions and Organization

The rest of this dissertation is organized as follows.

1.4.1 Background

Chapter 2 provides background information on the main topics discussed. The different building blocks of the SoC are introduced along with a summary of the main low power techniques they utilize. This chapter also presents an introduction to volatile and non-volatile memory cells with a description of their basic operation and their design challenges.

1.4.2 Low Voltage 6T SRAM

Chapter 3 focuses on reducing the operating voltage of a traditional six-transistor (6T) SRAM cell. This chapter investigates the use of combined read and write assist techniques to improve read and write functionality of SRAM cells at reduced voltages. The study presented in this chapter shows that while write failures initially limit voltage scaling, applying write assist introduces row and column half-select failures. Thus, read and write assist must be combined to allow voltage scaling down to sub-threshold voltages. We find that combining negative

bit-line (BL) for write assist with array V_{DD} boosting for read assist is most effective at minimizing the operating voltage and eliminating half-select failures. Knowledge of process corner also allows for further reduction through optimal control of the type and degree of assist applied.

1.4.3 Ultra-Low Power 8T SRAM

Chapter 4 presents a ULP 1KB SRAM array with a unique combination of features that makes it ideal for instruction memory in battery-less SoCs. These features allow the SoC to retain its program for a significantly longer period of time when energy harvesting conditions are poor. The array uses eight transistor (8T) low leakage SRAM cells with write assist to eliminate write failures, and a read-before-write scheme to address read-disturb in half-selected cells. To reduce the power consumption, a read burst mode is used when reading consecutive addresses, and aggressive power gating of all peripherals is employed during standby.

1.4.4 Low-V_{TH} STT-RAM

Chapter 5 investigates techniques to reduce the energy consumption of non-volatile STT-RAM arrays. Cell-level and array-level techniques are introduced to improve the write reliability of these cells while reducing their energy consumption. This chapter also describes a methodology for designing STT-RAM arrays that employ the proposed techniques.

1.4.5 Ferroelectric Auto-Recovery Sub-system (FeAR)

Chapter 6 describes a low power ferroelectric auto-recovery chip (FeAR) consisting of a ferroelectric non-volatile memory with a specialized ULP bus (ULP-BUS) interface to complement battery-less IoT SoCs. The proposed memory holds instructions and critical system data during power outages, and the ULP-BUS is designed to allow the integration of FeAR and the SoC in a compact System-In-Package (SiP).

1.4.6 Ultra-Low Power Always-On Voice Activity Detector

Chapter 7 introduces a sub-10nW always-on voice activity detector to be used as a wake-up for the self-powered SoC. The proposed detector uses a computationally inexpensive technique to detect voice activity in the presence of background noise. Digital techniques are used to reduce the required computation resources and thus the power consumption of this always-on block. Low power analog techniques are also utilized to ensure this always-on block remains within its power budget.

1.4.7 Sub-μW Self-powered SoC

Chapter 8 presents a completely self-powered sub-μW SoC. The proposed SoC has three sensing interfaces, a suite of hardware accelerators, a custom low power controller, a cold-boot and backup controller, an integrated energy harvesting and platform power manager, and an interface to a radio transmitter. The different components of the SoC are co-designed to enable a high level of integration while retaining ultra-low power operation. A cold-boot manager communicates with the auto-recovery sub-system to enable recovery from power failure. The platform power manager provides power to low power off-chip sensors that must interface with the SoC. It also provides power to the off-chip radio transmitter and the off-chip non-volatile memory.

1.4.8 Conclusion

Chapter 9 concludes this dissertation with a summary of the contributions and the broad impact of this work. This chapter also presents a list of interesting and challenging research questions that arise from this work.

Chapter 2

Background

This chapter introduces the main concepts discussed in this dissertation. Background information on the main blocks are presented here with a brief look at state-of-the-art techniques implemented in the literature. The chapter starts by presenting an overview of SoCs developed in the context of ULP or battery-less IoT devices: their main components and their low power techniques. Next, an overview of the basic operation, challenges and state-of-the-art techniques for SRAM, STT-RAM and Fe-RAM memories is presented.

2.1 IoT SoCs

2.1.1 Building Blocks

Figure 2.1 shows an abstract block diagram highlighting the common building blocks within an ULP SoC. Since the main function of such SoCs is to gather data, process it and then transmit it within a limited power budget, these SoCs usually include sensing interfaces, data processing units, communication interfaces, power monitoring or management units and clocking circuitry.

Sensing interfaces allow SoCs to gather important information that needs to be relayed. Thus, the exact nature of these interfaces depends hugely on the target application. Since

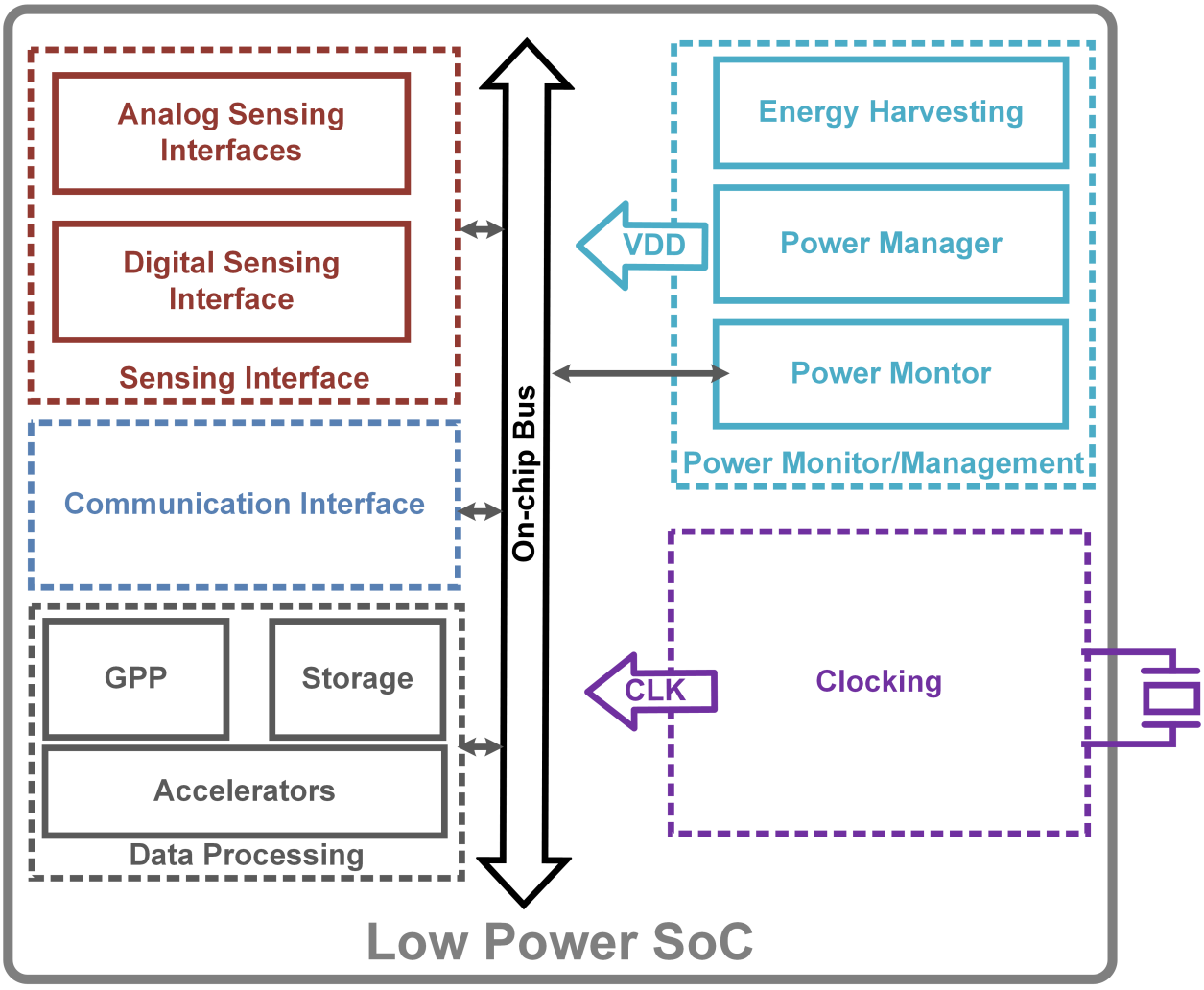


Figure 2.1: Block diagram showing the main building blocks of a typical low power SoC.

IoT devices are sufficiently diverse, typically one or more sensing interface are introduced. Some sensing interfaces are digital in nature such as wired communication interfaces (SPI, I2C, UART or simple wired inputs/outputs) to sensor chips that are difficult to integrate into the SoC. Other sensing interfaces are more complex and require specialized processing such as amplification, filtering and digitization. Such interfaces include ExG electrode inputs, microphone inputs and many others.

Once information is gathered from the sensors, **data processing** units collect this data and analyze it to detect important features. Data processing units include general purpose processors (GPP) that can handle any type of processing but are usually less efficient than

accelerators, accelerators that are especially efficient at performing particular tasks but cannot process any type of data, and memories to hold the required processing procedures and any generated data. Some SoCs rely on GPPs and do not include any accelerators and thus are flexible enough to target a wide variety of applications but might not be energy efficient for the target applications. Other SoCs include both a GPP and a number of accelerators that target a more specific application space and thus are more power and energy efficient for these applications. Nevertheless, all SoCs include storage elements of different sizes to accommodate the processing needs of the target application.

After the information is processed, **communication interfaces** are used to transfer this data to users. Typically, this information is sent to the users' smart phone, computer or even to the cloud. Thus, a wireless interface is required. However, in many ULP applications, the power consumption of wireless transceivers is significantly above the power budget of the device. To address this issue, the wireless transceiver is usually powered off when not used.

Since ULP SoCs have a limited power budget, they usually integrate a **power monitoring** unit or even a **power management** unit. The power monitoring unit usually keeps track of the available power to the system, and adjusts the system power consumption accordingly. On the other hand, the power management unit takes an unregulated supply and uses voltage regulators to provide stable supplies to the rest of the system at the required voltage(s). In self-powered systems, the power management unit usually includes an energy harvesting unit that boosts the input from a harvester and stores it on a super-capacitor or battery.

Finally, a **clock source** is always required in SoCs to enable time management and synchronous operation. An off-chip crystal oscillator typically interfaces to an on-chip oscillator with a phase-locked loop (PLL) to provide a stable clock source to the SoC. The PLL acts as a frequency multiplier that enables a wider frequency range depending on the requirements of the application.

Battery-less SoCs include many of these main components, and employ different power saving techniques to ensure that each component consumes as little power as possible. The

next section presents a survey of the available power saving techniques that are sometimes utilized by these systems.

2.1.2 Low Power Techniques

Many techniques were used in the literature to reduce energy and power consumption of circuits. These include: operating at near or sub-threshold voltages, dynamic voltage scaling, duty cycling, power and clock gating, using low-leakage devices, or using low-power logic families. In this section, we present a brief introduction to each of these techniques, their advantages and their limitations.

Due to the quadratic dependence of the active energy on the supply voltage, scaling the supply voltage of a circuit significantly reduces the active energy and power. However, as the voltage is scaled down to near or below the threshold voltage of the transistor, the leakage energy starts to dominate over the active energy [5]. This causes an increase in the total energy consumed to finish an operation, resulting in an optimal voltage that minimizes the total energy for that operation. Even though this technique reduces energy/power, it suffers from a number of adverse effects. The on current of the transistors at near/sub-threshold voltages is significantly reduced, and thus the on-to-off current ratio is degraded. This also results in a reduction in the maximum frequency at which the circuit can operate at.

While many low-throughput applications can handle the reduced performance, high-throughput applications cannot benefit from low-voltage operation directly. Applications with varying workload sometimes rely on dynamic voltage and frequency scaling (DVFS) to reduce the power/energy consumed to complete a low throughput task [6]. DVFS is used to decrease the slack time of a system and lower its throughput in the event an application's required performance is below what can be provided by the current voltage and frequency. Another approach that provides similar power savings with less overhead is voltage dithering [6] where only a few pairs of voltages and frequencies are implemented and are utilized alternatively to achieve intermediate voltage and frequency levels.

Duty cycling is another common power management technique. In this approach, high power circuits are powered intermittently in order to reduce the average power consumption when an application does not call for high performance. Duty cycling is often used with communication interfaces to reduce their contribution to the power budget [2][3]. Power gating and clock gating are methods used to implement duty cycling. Clock gating allows for a logic block to effectively pause operation. The clock signal is stopped from propagating at the input of a logic block when it is not used to reduce the cost of distributing it to the different registers within the block and reduce the blocks switching power/energy. Power gating completely cuts the power to a logic block to effectively cut down its power consumption when it is not used. To implement power gating, either PMOS headers are added between the power source and the logic block, or NMOS footers are added between the block and ground, or both headers and footers are added. Power gating is an effective power saving technique, however, it cannot be applied to storage blocks or blocks that require state retention. If the data contained within a logic block is not necessary to save, the block can be power gated, cutting off all power in order to save on leakage. When designing software that utilizes these blocks, these features can be used to aggressively decrease the power consumed by the system.

Another technique employed to reduce leakage power is the use of high threshold voltage (high-V_T) or low leakage transistors [7]. These devices usually have lower off current (I_{OFF}) than the standard devices. However, high-V_T transistors also have reduced on current (I_{ON}) making their use in the sub-threshold region particularly challenging. A mixture of devices with different threshold voltages can also be used to reduce the leakage power of the system. Different logic families styles [8][9] with lower leakage/active power were also introduced to replace standard CMOS. These styles rely on stacked logic and Schmitt trigger designs to reduce the leakage current of each gate.

2.2 SRAM

SRAM bit-cells are one of the major components of any system and one of the main contributors to its power consumption. Since SRAM bit-cells are volatile, they must be powered as long as their data is needed, even if the system is in a standby state. SRAM arrays are also large in size and suffer from variations that cause read, write and hold failures. This section will introduce SRAM bit-cells, describe their operation and the different challenges in their design, and present some of the state-of-the-art solutions to these challenges.

2.2.1 Basic Operation

The conventional 6T SRAM bit-cell shown in Figure 2.2 consists of two cross coupled inverters (PUL-PDL and PUR-PDR) that hold the data at nodes Q and QB, and two access transistors (PGL and PGR) used to read and write into the bit-cell. To read the bit-cell, the bit-lines (BL and BLB) are precharged to V_{DD} before enabling the word-line (WL). Assuming the bit-cell is holding "0" (on node Q), PGL will start discharging BL through PDL while PGR remains off, holding BLB at V_{DD}. Once a differential is developed on the bit-lines, a sense amplifier is enabled to sense and amplify the difference between the two bit-lines. To avoid accidentally writing into the bit-cell, PDL must be able to sink the current from BL faster than PGL can source it. This requirement forces SRAM designers to size the pull-down transistors (PDL and PDR) larger than the access transistors, thus allowing the pull-down transistors to sink more current.

During a write operation, one of the bit-lines is discharged according to the data that must be written, and then WL is asserted. Assuming the bit-cell originally holds a "1" and a "0" must be written, PGL will start discharging node Q while PGR starts to charge node QB. However, since PGR is weaker than PDR because of the read disturb requirement described above, PGL must sink node Q faster than PUL can source it. Thus, designers size the access transistors larger than the pull-up transistors (PUL and PUR).

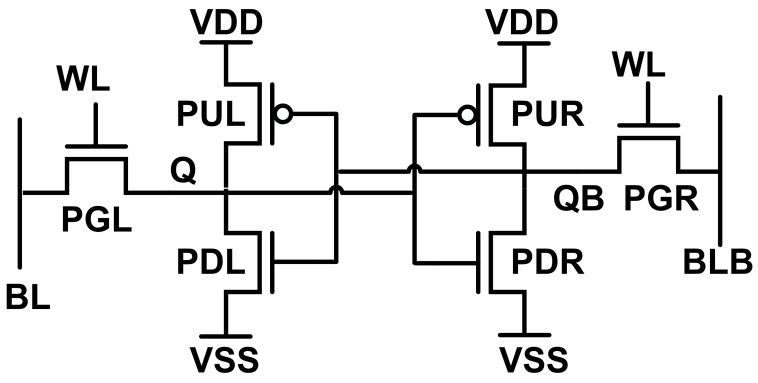


Figure 2.2: The Conventional 6T SRAM

2.2.2 Challenges

Since SRAM bit-cells rely on the relative strength of their transistors for correct operation, they are susceptible to variations in the manufacturing process. Variations are usually divided into two main categories: local and global variations. Local - or within-die (WID) - variations refer to threshold voltage variations between transistors on the same die, whereas global - or die-to-die (D2D) - variations refer to variations between transistors on different dice and are usually classified into five process corners. The threshold voltage of a transistor (V_T) can vary by as much as 230mV from one process corner to another. This wide variation in V_T of SRAM transistors significantly degrades its reliability. To aggravate the problem, a large number of SRAM bit-cells are required in any system which increases the probability of encountering high variations in the different transistors.

When the bit-cell is operated in sub-threshold, the impact of V_T variations becomes more pronounced due to the exponential dependence of the current on V_T in that region. Thus, variations in V_T might offset the impact of sizing and cause read, write and even hold failures. Also, the reduced on-to-off current ratio (I_{ON}-to-I_{OFF}) in sub-threshold might cause unselected bit-cells on the same column as a bit-cell being read to discharge a bit-line that is supposed to remain high resulting in incorrect reads. This significantly limits the number of cells that can be connected on the same bit-lines. To make matters worse, the drive strength of the pass-gate transistors might not be enough to create the required differential on the

bit-lines causing read failures.

2.2.3 Metrics

To account for all these challenges, SRAM designers measure different read, write and hold metrics to predict the behavior of the bit-cells under different variation profiles. These metrics can be divided into static and dynamic metrics. Static metrics determine the ability of the cell to perform the required operation assuming infinite time is available for it. Whereas, dynamic metrics determine the ability of the cell to perform the operation within a given time frame. Static metrics include the read (RSNM) and hold (HSNM) static noise margins [10] and the write margin (WM) [11]. To measure the HSNM and RSNM, a noise source is placed at one of the internal storage nodes of the SRAM bit-cell and its value is swept from 0 to V_{DD}. The voltage on the other node is determined and plotted to achieve the famous butterfly curves. The minimum between the diagonals of the two largest squares that can be fitted within the two eyes of the butterfly curve is the static noise margin. In the HSNM test, the access transistors are disabled; whereas in the RSNM test, the access transistors are enabled and the bit-lines are both precharged to V_{DD}. The WM is calculated by sweeping WL, measuring the value at which the contents of the cell switch, and then subtracting that value from V_{DD} [11].

Dynamic metrics include read and write delay and critical WL pulse. Read and write delays are defined as the amount of time required for the cell to perform a read/write operation. On the other hand, the critical WL pulse is defined as the minimum time WL must be held high for the cell to complete its operation successfully. While the two metrics might sound the same, it has been observed that the WL does not need to remain on until the completion of the operation for it to be successful. Thus, many high performance SRAM designers rely on and optimize the critical WL pulse to determine and improve the maximum frequency to operate at without failure.

2.2.4 Low Power Techniques for SRAMs

Many approaches were introduced in the literature to address the different challenges facing sub-threshold SRAM operation. A number of alternative bit-cell topologies were used in [12][13][14] including the 8T bit-cell [12] shown in Figure 2.3 with decouple read and write ports to improve the stability for sub-threshold operation. Different assist techniques [15] were also used to improve read and write stability at lower supply voltages. Assist techniques modify the strengths of different transistors in the cell by either boosting or suppressing their gate-to-source voltages. However, introducing write assist techniques induces stability problems in half-selected cells that experience pseudo-read during a write operation. To eliminate half-select instability, new bit-cell topologies [16], banks with only one word per row [13] [16], and row read-before-write [17] were proposed. Different combinations of read and write assist techniques were also used to address this problem [18][19]. The authors in [18] suggested combining either negative BL (NegBL) or lowered column V_{DD} (LC_{VDD}) with under-driving WL (UDWL) to eliminate half-select failures while maintaining write stability and speed. In [19], NegBL was combined with UDWL to improve both read and write stability.

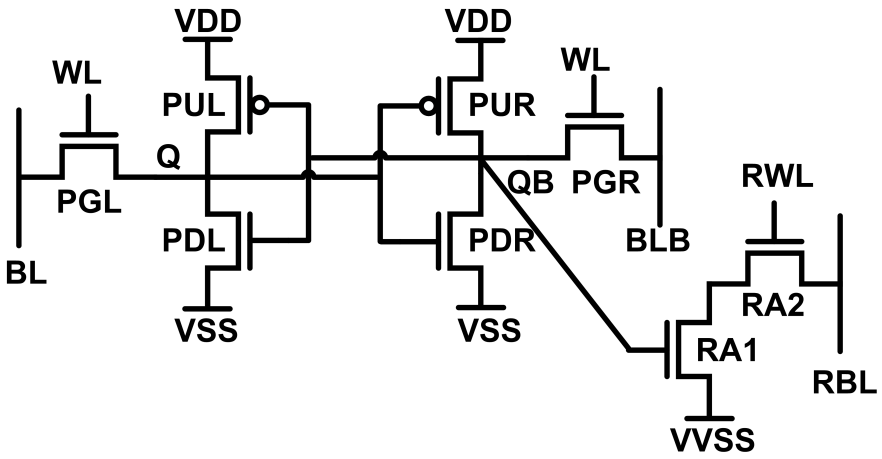
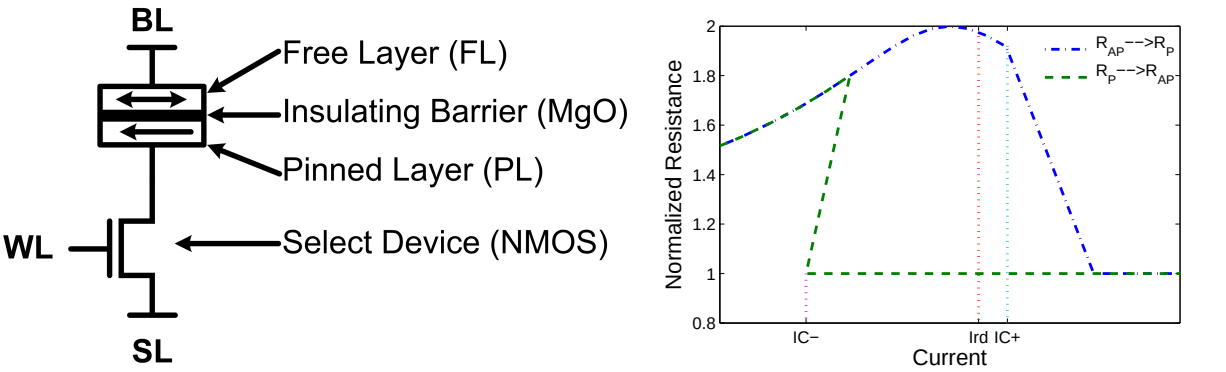


Figure 2.3: The Conventional 8T SRAM [12]

2.3 STT-RAM

STT-RAM is a type of emerging non-volatile memory cell in which a bit is stored on a Magnetic Tunnel Junction (MTJ) and accessed through a select device usually an NMOS transistor (Figure 2.4a). The MTJ is formed of two ferromagnetic layers separated by an insulating layer. One of the layers - the pinned layer (PL) - has a fixed spin orientation, whereas the orientation of the other layer - the free layer (FL) - determines the stored data. When the spins of the two layers are parallel (P), the MTJ has low resistance (R_P) and is considered to hold a "0". On the other hand, when the spins are antiparallel (AP), the MTJ has high resistance (R_{AP}) and is considered to hold a "1". Figure 2.4b shows how the resistance of the MTJ changes when the current passing through it is varied.

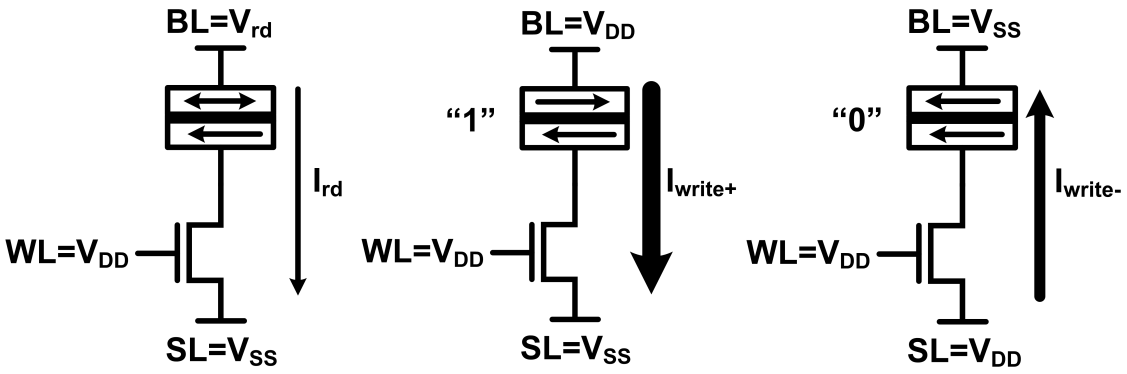


(a) STT-RAM bit-cell (b) MTJ resistance vs. current characteristic

Figure 2.4: STT-RAM bit-cell and MTJ characteristics

2.3.1 Basic Operation

To read the STT-RAM cell (Figure 2.5a), a small current (I_{rd}) is supplied to the cell and the resulting voltage (V_{rd}) is measured and compared to a reference [20]. Two types of errors can occur during a read operation: *read distinguishability* errors and *read disturb* errors [20]. If the current passing through the MTJ is not enough to create the corresponding V_{rd} required to correctly read the data, a distinguishability error has occurred. On the other hand, if the current passing through the MTJ is high, then an accidental write might occur and this is



(a) Read Operation (b) Write "0" (c) Write "1"

Figure 2.5: STT-RAM read and write operations

known as a read disturb error. Thus, the choice of read current and reference voltage are crucial to the success of the read operation.

To write "0" into the STT-RAM cell (Figure 2.5b), current must be passed from the BL to the SL allowing electrons of both spin to reach the PL. The PL then acts as a spin filter allowing electrons with its spin orientation to continue to the FL. Those electrons will then exert a torque on the FL forcing it to orient its spin parallel to the PL when the current through it exceeds the critical AP-to-P switching current (I_{C+}). During a write "1" operation (Figure 2.5c), current flows in the opposite direction. However, the FL is not a good spin filter and thus electrons of both spin directions will reach the PL which will then reflect the electrons with opposite spin back into the FL where they exert a torque forcing the FL to reverse its spin orientation to become anti-parallel to the PL when the current passing through it exceeds the critical P-to-AP switching current (I_{C-}) [21]. The mechanism for switching from "0" to "1" is therefore more difficult and requires a larger current to complete ($I_{C-} > I_{C+}$). To make matters worse, the NMOS is in a source follower (SF) configuration and thus is not able to supply a high current to switch the device. With further scaled technologies, the interconnect resistance will also reduce the current supplied to the cell and make the write operation even more challenging. Another issue with the write operation is that the large current required to switch the MTJ will cause a high voltage across it which might cause the oxide barrier to break down if the MTJ voltage exceeds the breakdown

voltage (V_{BD}) [22]. The V_{BD} of different MTJs within the array has a distribution [22] and thus the V_{BD} limit chosen in our analysis is the least possible value in that distribution.

2.3.2 Challenges and Available Solutions

One important problem impacting the scaling of STT-RAM cell is the high spin current required to switch the MTJ. This critical write current is in direct trade off to 1) the area of the MTJ device [20], 2) write speed [20], 3) write energy [20], 4) MTJ oxide (MgO) break-down limit (V_{BD} violation) [22], and 5) the non-volatility of the MTJ, commonly measured by its thermal stability [22]. Further adverse effects are manifested by die-to-die (D2D) and within-die (WID) variations in both the MTJ and the select transistor, especially when designing for large arrays requiring margining for $6\sigma+$ of variations to meet high yield target.

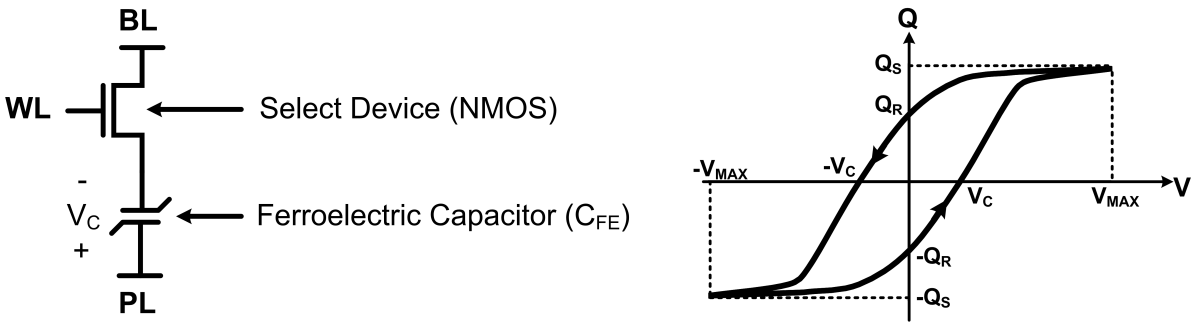
Switching the MTJ from P-to-AP configuration requires a large write current and consumes significant write energy. To supply the large current, the size of the NMOS is usually increased or WL is boosted. Increasing the size of the NMOS increases the current through the MTJ at the expense of cell area, V_{BD} violation risk, and unnecessarily higher power for other write cases (P-to-P, AP-to-AP and AP-to-P). Boosting the WL voltage beyond V_{MAX} can also increase the current through the NMOS during write operation but can cause NMOS gate oxide breakdown and requires a charge pump.

The authors in [23] proposed using a negative BL voltage during a write "1" operation (while SL is grounded) to reduce the SF impact. Besides requiring a negative rail generator, this approach requires under-driving the WL gate voltage below V_{DD} and tying the NMOS body to the negative rail. Thus, this approach requires triple-well technology and an additional WL under-drive generator. In [24], the authors proposed reducing the write energy consumption by terminating the write operation early if the same data is being written. The authors in [25] proposed using perpendicular magnetic anisotropy (PMA) MTJ instead of the commonly used in-plane MTJ since it provides more balanced writes but the manufacturing

of these MTJs is not as advanced as in-plane MTJs. In [26], the authors use a PMA MTJ structure to study write ability of STT-RAM in an array setup. The authors also raise the SL voltage above V_{SS} during a write "0" operation and provide a WL voltage greater than V_{DD} to ensure correct operation. Neither the motive nor the means for raising the SL were stated in that work.

2.4 Fe-RAM

Ferroelectric Random Access Memory - FeRAM - is another class of promising low power CMOS-compatible non-volatile memories. In FeRAMs, a bit of data is saved on to a ferroelectric capacitor and accessed through a select transistor (Figure 2.6a). Ferroelectric capacitors resemble normal capacitors but use ferroelectric materials as dielectrics. Like conventional dielectric materials, ferroelectric materials experience polarization when an electric field is applied. However, these materials retain their polarization even after the electric field is removed. This polarization can be reversed by applying an electric field of the opposite direction across the ferroelectic capacitor. Thus, the direction of the polarization is used to represent the data to be saved within the ferroelectric capacitor. Figure 2.6b shows the change in the charge on the ferroelectric capacitor as a function of the voltage applied across it. A positive charge within the capacitor represents a "0"; whereas, a negative charge represents a "1".



(a) FeRAM bit-cell

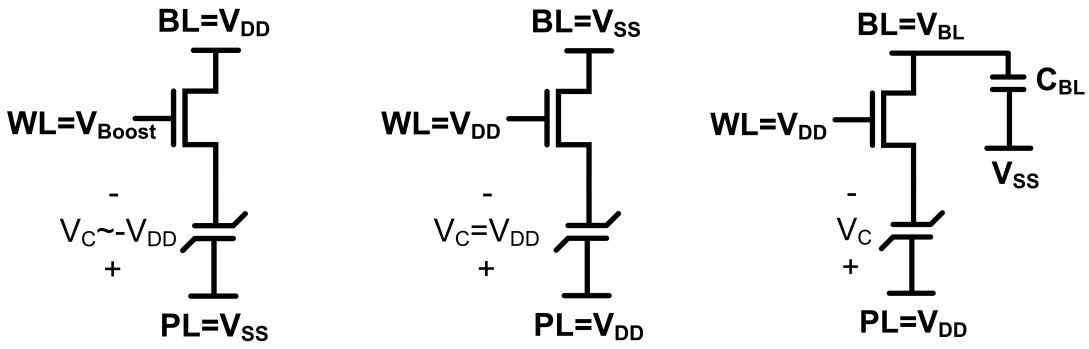
(b) Charge vs. voltage characteristic [27]

Figure 2.6: FeRAM bit-cell and ferroelectric capacitor characteristics

2.4.1 Basic Operation

To write a "1" into the FeRAM cell (Figure 2.7a), BL is driven to V_{DD} while PL is grounded before WL is enabled. Thus, a negative voltage (V_C) is applied across the capacitor forcing it to become negatively charged. WL is usually boosted to guarantee a full swing voltage across the capacitor ($V_C = -V_{DD}$). To write a "0" into the FeRAM cell (Figure 2.7b), PL is driven to V_{DD} while BL is grounded before WL is enabled. In this case, a positive voltage appears across the capacitor forcing it to become positively charged. Traditionally, WL and PL are shared among cells in the same row [27]. Thus, to write different data simultaneously to different cells in the same row, BL of each cell is first driven to the correct value, then WL is driven high. PL is then kept low for some time before it is driven high. If a cell is experiencing a write "0" operation, BL is kept low, thus when PL is low, the cell contents do not change. When PL is driven high, the capacitor will experience a positive voltage across it and change its polarization. Similarly, if a cell is experiencing a write "1" operation, BL is driven high, thus when PL is low, a negative voltage will be applied to its capacitor forcing a negative charge into it. When PL is driven high, the voltage across the cell will be zero and thus it will hold its state.

To read the FeRAM cell, BL is first discharged to V_{SS} then left floating. Next, PL is driven high and WL is enabled. The charge within the FeRAM cell will be shared with the capacitance on BL (C_{BL}) and a voltage is developed on BL. A sense amplifier is then used to compare the developed voltage (V_{BL}) to a reference voltage (V_{REF}), and to read out the data. Since reading a "1" will corrupt the data saved within the FeRAM cell, WL remains enabled for a short period after the sense amplifier has read out the data. During this period, if the cell was holding "1", BL will be driven high by the sense amplifier. PL is then driven low to re-write "1" into the cell before WL is disabled. If the cell was holding a "0" instead, BL will be driven low by the sense amplifier, thus when PL is driven low, the cell will experience zero voltage across it and the capacitor will remain undisturbed.



(a) Write "1" (b) Write "0" (c) Read

Figure 2.7: FeRAM read and write operations

2.4.2 Challenges and Available Solutions

Fe-RAMs face a number of challenges that limit their widespread use in battery-less SoCs. Like all non-volatile memories, Fe-RAMs consume significantly higher read and write energy than SRAM (∼31pJ/bit [27] compared to ∼1.7fJ/bit). Reading an Fe-RAM disturbs its contents requiring a write back operation. An accurate reference voltage is also needed to reliably read the cell contents but is hard to achieve since it must be generated using dummy Fe-RAM cells. Imperfections in the ferroelectric capacitors aggravate the read challenges by reducing the differential voltage between "0" and "1" [27]. Imprints occur when the contents of the ferroelectic capacitor is kept constant for an extended period of time, causing less differential between "0" and "1" for that cell [27].

Many techniques were introduced in the literature to address the read challenges facing Fe-RAMs. These techniques either target the bit-cell, the reference generators, the array architecture or the sensing scheme [28]. Bit-cell techniques rely on eliminating the need for a reference generator by saving the data and its complement either on two [29] or four [30] ferroelectric capacitors. While this approach improves the reliability of the read operation, it increases the area requirement of the Fe-RAM array. Array architecture solutions usually introduce either a row or a column of reference cell. However, row-based reference cells experience higher access rate than normal cells and thus are subject to fatigue which reduces their accuracy [27]. Column-based reference cells require complex sensing circuitry to

eliminate the extra capacitance required to route the reference line to every sense-amplifier [27]. Thus, generating an accurate reference for Fe-RAM cells remains a challenge.

Chapter 3

Low Voltage 6T SRAM

[1] Section 1.1 identified SRAM bit-cells as one of the major components of any system and one of the main contributors to its power consumption. One of the most effective techniques to reduce the power consumption of SRAM arrays is scaling V_{DD} to near/sub-threshold voltages. However, operating SRAM bit-cells at such a reduced voltage has significant challenges. This chapter evaluates the different read and write assist techniques. Even though assist technique were evaluated in the literature [31][32], we study the effects of assist on both the write stability of a selected bit-cell and the read/hold stability of row and column half-selected cells for a wider supply voltage range, and use sensitivity analysis [33] to explain the unique results observed at lower supply voltages. Then, we propose to simultaneously apply a write assist technique to improve the write-ability and a read assist technique to reduce half-select failures. The chapter ends by showing the advantages of employing a process monitor to control the degree of assist applied.

3.1 Write Assist Evaluation

First, we evaluated the impact of the three most commonly used write assist techniques – negative BL (NegBL), WL boosting (BWL) and column V_{DD} lowering (LCV_{DD}) – on the

[1] This chapter is based on [FBY1][FBY2]

write, read and hold stability of the 6T bit-cell. The static write margin (WM) [11], read and hold static noise margins (RSNM/HSNM) [10] are used as metrics to quantify the write, read and hold stability, respectively. Monte-Carlo simulations are performed to determine the worst case WM at the SF process corner with a temperature (T) of 25^0C since that corresponds to the worst case write corner. The worst case RSNM and HSNM are measured at the worst case half-select corner (FS 100^0C). The assist voltages are chosen as different percentages of the applied V_{DD}.

Figure 3.1 shows the impact of different write assist techniques on WM. High percentages of assist (40%) are required to reduce the write V_{MIN} of the accessed bit-cell down to sub-threshold voltages. Introducing (40%) NegBL lowers the write V_{MIN} down to 450mV, while BWL and LCV_{DD} reduce it further down to 400mV. BWL shows the most improvement in WM at V_{DD} higher than 500mV, whereas LCV_{DD} shows the most improvement at lower V_{DD}s.

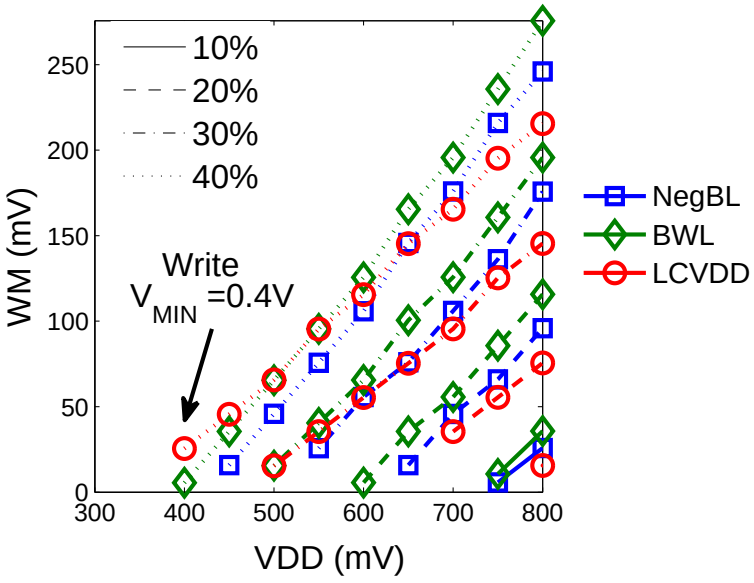


Figure 3.1: Impact of write assist on WM at the worst case write corner (SF 25^0C).

To explain the change in the most effective assist with V_{DD}, Figure 3.2 shows the sensitivity of WM to changes in the V_T of each of the six transistors. WM is very sensitive to changes in the V_T of PGR (Figure 2.2). PUR, PDL, and PGL also impact WM, while PUL and PDR

have negligible impact on WM. Since BWL improves the strength of both PGR and PGL, it gives the highest WM at high V_{DD}s. NegBL improves the strength of PGR only and thus gives a lower WM than BWL. Even though LCV_{DD} reduces the strength of PUR, at high V_{DD}s, WM is more sensitive to changes in PGR, and thus LCV_{DD} gives lower WM than BWL and NegBL. As V_{DD} is scaled however, the sensitivity of WM to changes in PUR and PDL increases, and thus LCV_{DD} becomes more effective at lower V_{DD}s, since it reduces the strength of both PUR and PDL.

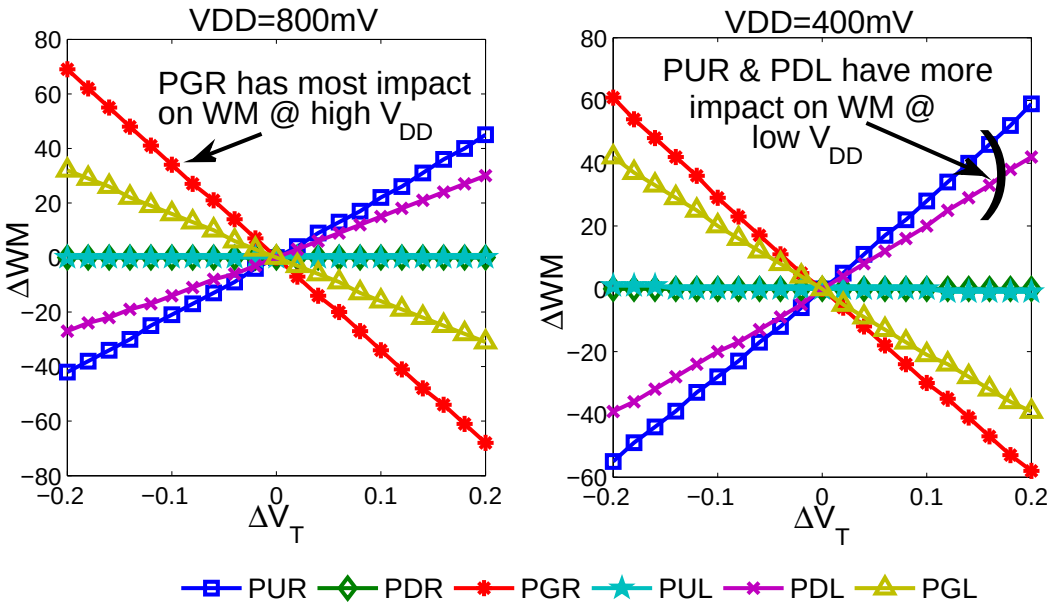


Figure 3.2: Sensitivity of WM to changes in V_T of the SRAM transistors at different V_{DD} values.

Even though the different assist techniques reduce the write V_{MIN}, they negatively impact the SNM of row and column half-selected bit-cells, which share the WL and BLs of the selected cells, respectively. Both LCV_{DD} and NegBL reduce the HSNM of column half-selected bit-cells but have no impact on the RSNM of the row half-selected bit-cells. BWL, on the other hand, has no impact on the HSNM of column half-selected bit-cells but lowers the RSNM of row half-selected bit-cells, causing the half-select V_{MIN} to increase as shown in Figure 3.3. Applying 40% BWL increases the row half-select V_{MIN} from 700mV to above

800mV, while applying 40% LCV_{DD} or NegBL degrades the HSNM and raises the column half-select V_{MIN} from 350mV to 750mV or above 800mV, respectively.

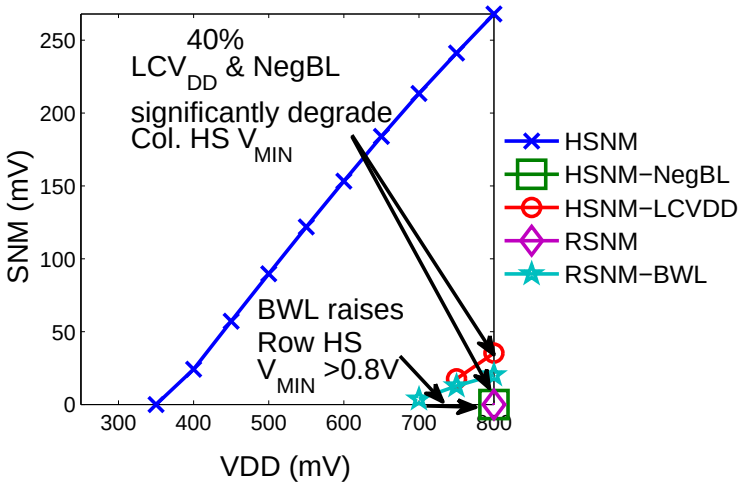


Figure 3.3: Impact of applying 40% write assist on half-selected bit-cells at the worst case half-select corner (FS 100^0C)

To understand the impact of write assist on the RSNM and HSNM, Figure 3.4 and Figure 3.5 show the sensitivity of these margins to changes in the V_T of the SRAM transistors. At high V_{DD}s, the RSNM is most sensitive to PGL and PDR (Figure 2.2). PUR and PDL also impact RSNM while PUL and PGR have negligible impact. At lower V_{DD}s, PUR and PDL have a higher impact on RSNM. However, when the V_T of PUR is above a certain limit, its impact on RSNM becomes fixed. When PGR is strong (lower V_T), it helps improve RSNM since BLB is high and leakage currents through PGR help retain the data. This trend is more obvious at lower V_{DD}s since the I_{ON}-to-I_{OFF} ratio is lower and thus leakage current through PGR is comparable to current through PUR. BWL improves the strength of PGL, causing the node holding "0" to rise, thus reducing the strength of PUR and improving the strength of PDR. This effect reduces the RSNM, causing the number of failures and the half-select V_{MIN} to increase.

According to Figure 3.5, the HSNM is most sensitive to PUR and PDR at high V_{DD}s. However, at lower V_{DD}s, the impact of PUR increases. When PGR is strong, it negatively

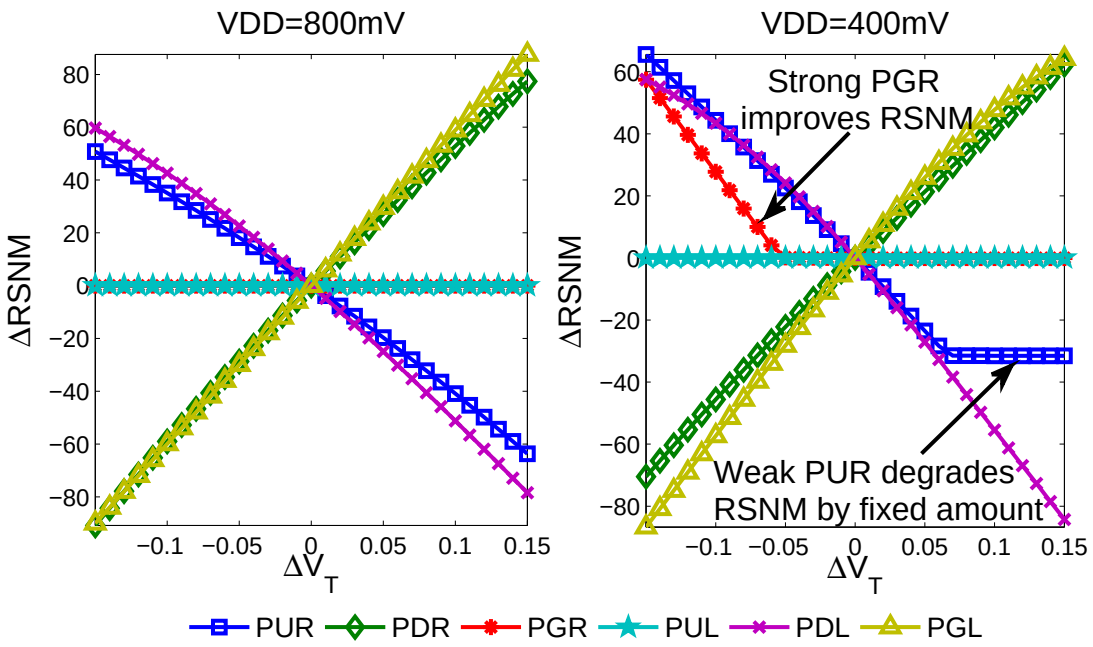


Figure 3.4: Sensitivity of RSNM to changes in V_T of the SRAM transistors at different V_{DD} values.

impacts the HSNM since the leakage current through it is comparable to the on-current through PUR and the fight between the two transistors will determine the voltage at the node holding "1". Applying 40% NegBL on BLB will increase the strength of PGR, turning it partially on and causing the bit-cell to lose its data even at high V_{DD} values. Applying 40% $\mathrm{LCV_{DD}}$ degrades the strength of PUR and thus increases the half-select V_{MIN} up to 750mV. At higher V_{DD}s, the HSNM is less sensitive to PUR changes. This causes the HSNM to reduce significantly but remain positive.

To reduce the write V_{MIN} down to sub-threshold voltages, 40% write assist must be applied. Table 3.1 summarizes the impact of write assist on the write and half-select V_{MIN}. Even though write assist techniques reduce the write V_{MIN}, they increase the HS V_{MIN}, thus limiting the array level V_{MIN}. Column based write assist techniques ($\mathrm{LCV_{DD}}$ and NegBL) degrade the hold stability (HSNM) of column half-select bit-cells and have no impact on the row half-select bit-cells. On the other hand, row based write assist techniques (BWL) have no impact on the hold stability of column half-select bit-cells but significantly degrade the read stability (RSNM) of row half-select bit-cells. In both cases, write assist techniques

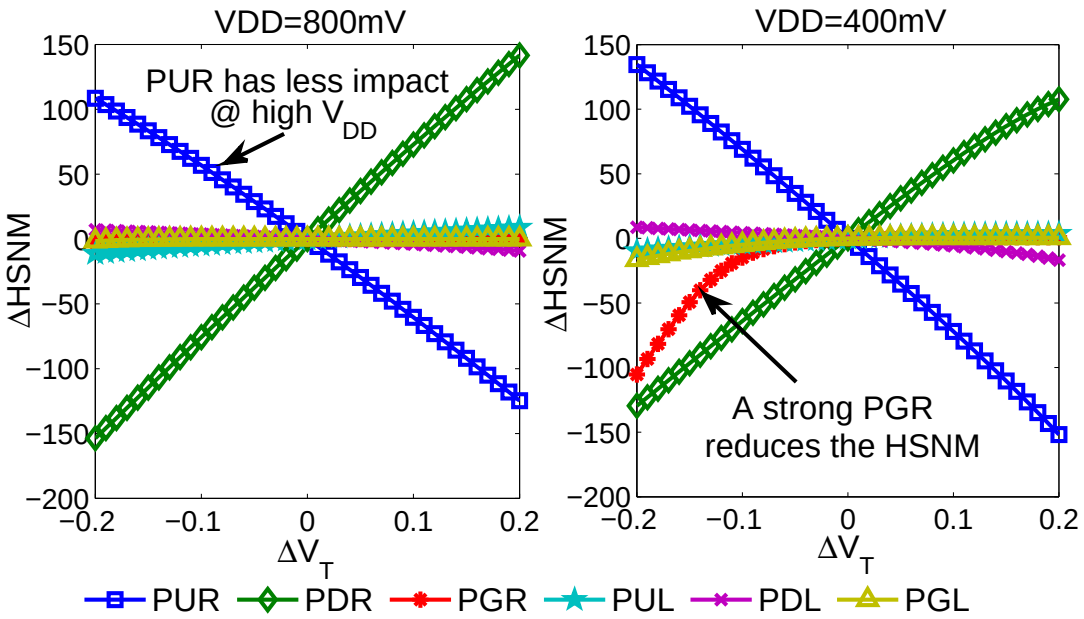


Figure 3.5: Sensitivity of HSNM to changes in V_T of the SRAM transistors at different V_{DD} values.

reduce the write V_{MIN} below the half-select V_{MIN} making half-select failures the limiting factor on the array V_{MIN}.

Table 3.1: Write and half-select V_{MIN} with 40% applied write assist in mV.

	Write V_{MIN}	**Half-select V_{MIN}**
No Assist	>800	700
NegBL	450	>800
LCV_{DD}	400	750
BWL	400	>800

3.2 Read Assist Evaluation

So far, we determined that half-select failures will limit the array V_{MIN} when write assist techniques are employed. Now, we evaluate the impact of read assist techniques - V_{DD} boosting (RV_{DD}) and WL under-driving (UDWL) - on the write and half-select V_{MIN}. Monte-Carlo simulations at different process corners show that the FS corner is the worst case corner

for the half-select bit-cells. Thus, the FS corner with temperature set at 100^0C is used to measure the worst case RSNM and HSNM.

Figure 3.6 shows the impact of read assist techniques on the worst case RSNM at different V_{DD}s. Without any read assist, the row half-selected bit-cells experience a read upset for V_{DD}s below 700mV. Applying 20% UDWL improves the RSNM of these bit-cells and enables correct functionality down to 500mV. Applying 20% $\mathrm{RV_{DD}}$ provides additional improvements in the RSNM and lowers the half-select V_{MIN} further down to 450mV.

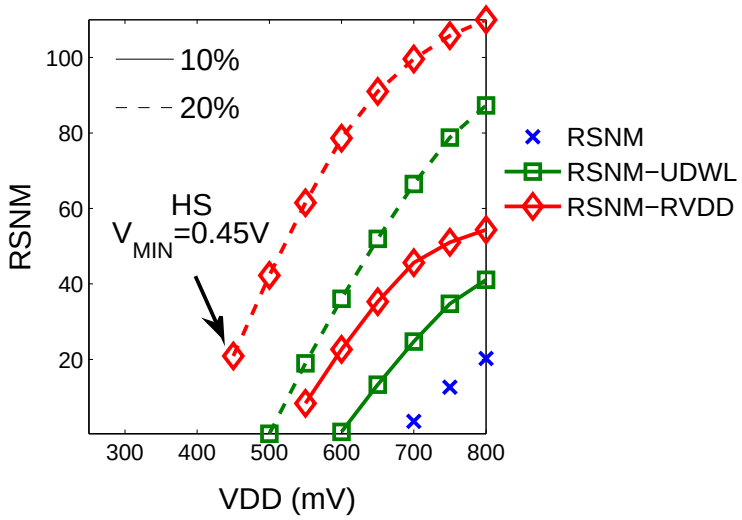


Figure 3.6: Impact of read assist on RSNM at the worst case half-select corner (FS 100^0C)

$\mathrm{RV_{DD}}$ shows more improvement in the RSNM than UDWL. Referring back to Figure 3.4, UDWL reduces the strength of PGL, which means that the voltage on node "0" does not rise significantly, and thus PDR is kept firmly off. On the other hand, $\mathrm{RV_{DD}}$ improves the strength of PUR and PDL (since the gate of PDL is driven by a higher bias). A stronger PDL also means node "0" remains low keeping PDR firmly off. At lower V_{DD}s, the sensitivity of RSNM to changes in PUR and PDL increases making $\mathrm{RV_{DD}}$ more effective at improving the RSNM and reducing the half-select V_{MIN}.

Read assist techniques improve the read stability of row half-selected bit-cells at the cost of degraded write stability in selected bit-cells. The accessed bit-cell fails to write without

write assist techniques; applying read assist techniques will only increase the number of failing bit-cells within the array at a particular voltage. Table 3.2 shows the impact of applying these percentages of assist on the write and half-select V_{MIN}.

Table 3.2: Write and half-select V_{MIN} with 20% applied read assist in mV.

	Write V_{MIN}	**Half-select V_{MIN}**
No Assist	>800	700
RV_{DD}	>800	450
UDWL	>800	500

3.3 Proposed Read/Write Assist Combination

Using read and write assist techniques independently does not reduce the overall array V_{MIN}. Thus, we propose combining NegBL (write assist) and array RV_{DD} (read assist) to achieve a lower array V_{MIN}. Array V_{DD} is defined as the V_{DD} of the SRAM bit-cell array only, excluding drivers and peripherals. After evaluating the read and write assist techniques independently, we now evaluate the proposed combination of RV_{DD}–NegBL and compare it to the two previously proposed combinations [18][19]. We define UDWL-NegBL as the combination of UDWL (read assist) and NegBL (write assist) proposed in [18][19], and UDWL-LCV_{DD} as the combination of UDWL (read assist) and LCV_{DD} (write assist) proposed in [18].

Since the RSNM and HSNM are pessimistic measures of half-select failure, we look at dynamic half-select failures by running transient simulations with a relaxed WL pulse width. Figure 3.7 shows the impact of applying different percentages of NegBL on the write delay. Column half-select failures limit the percentage of NegBL that could be applied to below 40%. Applying 30% NegBL reduces the write V_{MIN} from above 800mV to 550mV, however, row half-select failures limit the array V_{MIN} to 650mV.

To address the row half-select failures, 10% or 20% UDWL is applied simultaneously with 30% NegBL. Since UDWL does not reduce the column half-select failures, a higher percentage of NegBL cannot be used. Figure 3.8 shows the impact of combining 30% NegBL

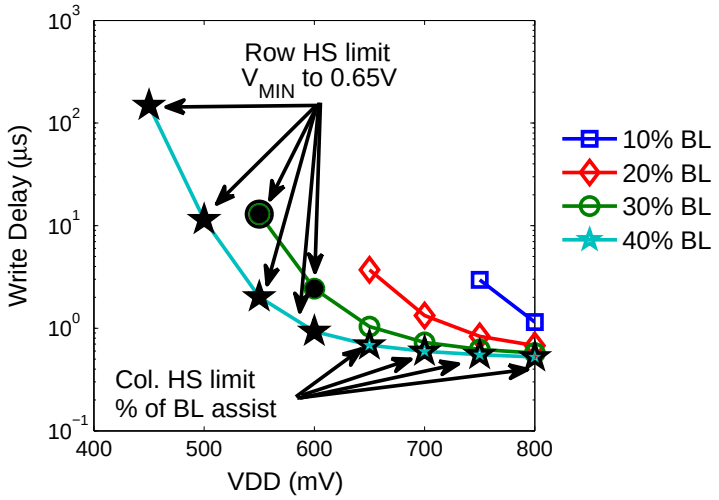


Figure 3.7: Impact of NegBL on the write delay - black markers indicate half-select (HS) failures.

with different percentages of UDWL. Even though UDWL reduces the row half-select V_{MIN}, it increases the number of write failures and raises the write V_{MIN}. Combining 10% UDWL with 30% NegBL does not change the array V_{MIN}. Reducing the WL further (20% UDWL) will significantly degrade the write V_{MIN} resulting in an overall degradation in the array V_{MIN}. Thus, UDWL-NegBL does not provide any advantages in V_{MIN} over using only NegBL.

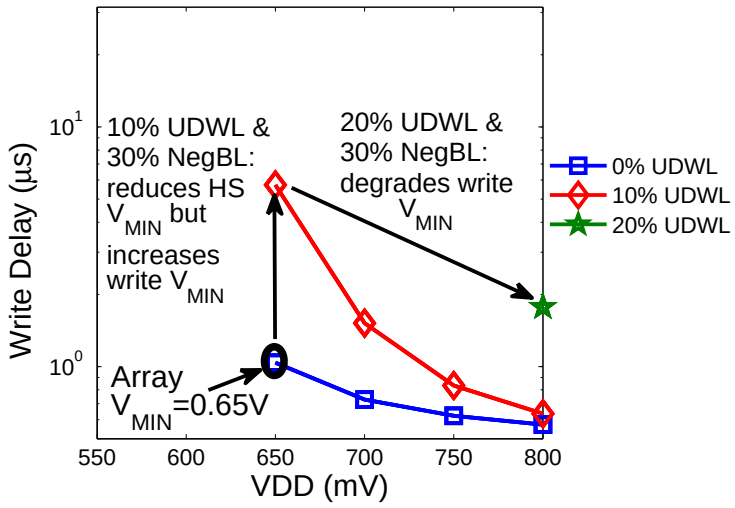


Figure 3.8: Impact of UDWL–NegBL on the write delay.

UDWL can also be combined with LCV_{DD} to reduce the overall array V_{MIN}. Figure 3.9

shows the impact of applying 30% LCV_{DD} with UDWL. While this combination reduces the half-select and write V_{MIN}, developing enough BL/BLB differential to perform a correct read operation is not possible with the UDWL (black marker in Figure 3.9). Thus, differential read V_{MIN} limits the overall V_{DD} to 650mV.

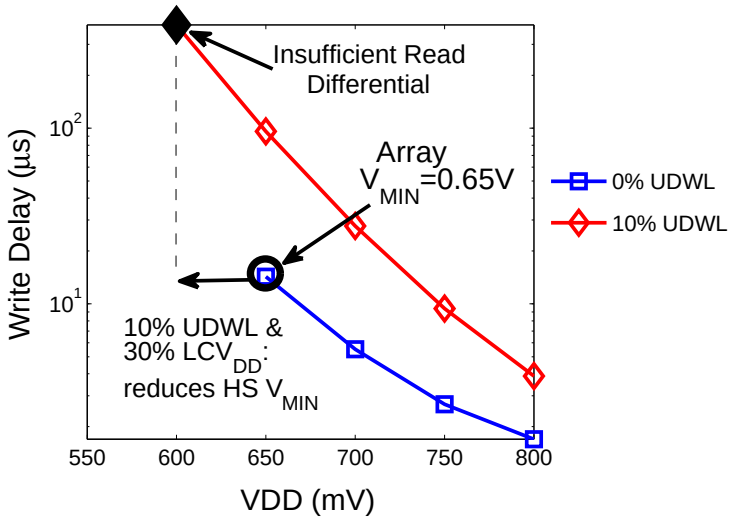


Figure 3.9: Impact of UDWL–LCV_{DD} on the write delay.

The proposed RV_{DD}-NegBL improves V_{MIN} most effectively. Boosting the array V_{DD} helps address both row and column half-select failures allowing a lower NegBL to be used if needed. It also improves the read capabilities of the bit-cell and ensures enough BL/BLB differential is developed for a correct read operation. Figure 3.10 shows the impact of combining 30% NegBL with different percentages of array RV_{DD}. 10% RV_{DD} is enough to reduce the row half-select failures and V_{MIN} down to 600mV. 20% RV_{DD} increases the write V_{MIN} (to 700mV) and reduces the half-select failures but does not allow a higher percentage of NegBL assist. 30% RV_{DD} improves the column half-select V_{MIN} allowing 40% NegBL to be used. However, this combination does not reduce the array V_{MIN} below 650mV. Even though a large percentage of NegBL is required to push V_{MIN}, the drive voltage across the access transistors and within the NegBL generation circuit does not exceed the V_{MAX} of the technology, thus oxide breakdown is not a concern.

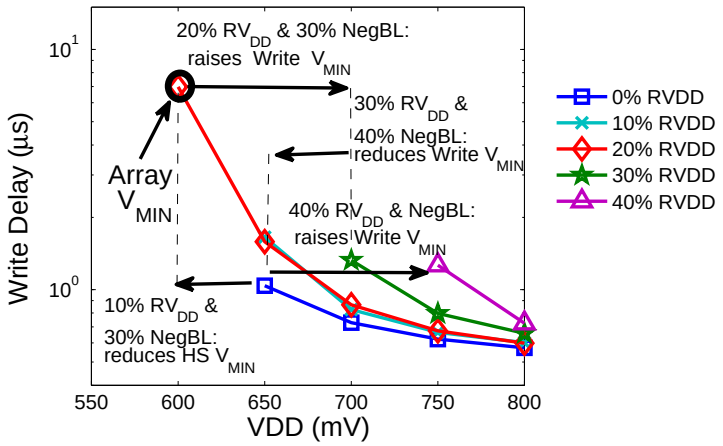


Figure 3.10: Impact of RV_{DD}-NegBL on the write delay.

Table 3.3 provides a comparison between the proposed combination (RV_{DD}-NegBL) and the combinations in [18] and [19]. RV_{DD}-NegBL is the most effective combination for reducing the array V_{MIN} to near/sub-threshold voltages. The combination of read and write assist techniques provide significant improvement in V_{MIN} compared to the V_{MIN} achieved without any assist applied. The proposed combination provides more than 25% improvement in V_{MIN} over the no assist case. Compared to the most effective write assist technique (BWL from Table 3.1), the proposed combination provides more than 20% improvement in the V_{MIN}.

Table 3.3: V_{MIN} comparison between the proposed assist combination and previous state-of-the-art combinations.

	Array V_{MIN} (mV)	**% lower V_{MIN} vs. No Assist**	**% lower V_{MIN} vs. Write Assist only**
Proposed	600	>25%	>20%
[18][19]	650	>19%	>13%
[18]	650	>19%	>13%

3.4 Corner Analysis

From the evaluation in the previous section, we noticed that the worst case write corner (SF) differs from the worst case half-select corner (FS). Thus, if a process monitor is available

within the system, knowledge of the process corner can help in reducing V_{MIN} and the degree of assist needed to achieve this V_{MIN}. To determine the effectiveness of this technique, we ran Monte-Carlo simulations for the proposed combination of read/write assist (RV_{DD}-NegBL) at each process corner at two temperatures (25^0C and 100^0C) to determine the required assist percentages to reduce V_{MIN}.

Table 3.4 shows the percentages of NegBL and array RV_{DD} required to achieve a V_{MIN} of 450mV. For the TT and SS, applying 10% NegBL is enough to guarantee correct write operation without disturbing row and column half-selected cells down to 450mV. At the FF corner, write assist is not required down to 450mV but 10% RV_{DD} is needed to address half-select failures at this V_{MIN}. For the worst case write corner (SF), only 40% NegBL is needed since at this corner row and column half-selected cells are not disturbed. At the worst case half-select corner (FS), 20% RV_{DD} will eliminate half-select disturbances and no write assist is required. Thus, considering the process corner of the chip allows us to achieve a much lower array V_{MIN} and also reduces the required percentage of assist needed to achieve this V_{MIN}.

Table 3.4: Required assist per corner to reduce array V_{MIN} to 450mV

Corner	**NegBL**	RV_{DD}	V_{MIN} **(mV)**
TT	10%	0%	450
SS	10%	0%	450
FF	0%	10%	450
SF	40%	0%	450
FS	0%	20%	450

3.5 Conclusion

In this chapter, a combination of read and write assist techniques was introduced to reduce the V_{MIN} of SRAM arrays down to near/sub-threshold voltages while addressing row and column half-select failures. A detailed evaluation of the read and write assist techniques for a

wide range of supply voltages was also presented. The proposed combination - array RV_{DD} and NegBL - allows maximum reduction in the array V_{MIN} (to 600mV) when compared to other combinations available in the literature [18] [19]. This chapter also presented the advantages of employing a process monitor to control the percentages of assist applied in order to reduce the overall array V_{MIN}. Controlling the applied assist per corner allows us to reduce the array V_{MIN} to 450mV while only applying one type of assist at each corner (NegBL for TT, SS and SF; and RV_{DD} for FF and FS).

Chapter 4

Ultra-Low Power 8T SRAM

[1] The previous chapter looked into techniques to reduce the minimum operating voltage of the 6T SRAM bit-cell. However, using this bit-cell, the minimum achievable V_{MIN} is 450mV. In this chapter, we investigate the 8T bit-cell, and build a ULP 1KB SRAM array capable of operating reliably down to 350mV. The proposed array is designed for battery-less SoCs, and is used as a building block for two different types of memories: a "read-mostly" memory to hold the program instructions that run on the SoC, and a "read-write" memory to save the data gathered. Many of the design decisions and added features in this array aim at reducing its power and energy consumption. Table 4.1 summarizes these features.

4.1 Array Structure

Figure 4.1 shows the overall structure of the array. The 1 KB array consists of 64x128 8T bit-cells with row (RDx) and column (CDx) drivers, a row decoder, a read/write control unit with a burst control unit (BCU), and a data management unit (DMU). In the next subsections, the main features of each unit are described.

[1] This chapter is based on [FBY4][FBY7]

Table 4.1: Main features of the ULP 1KB SRAM chip

Technology	Commercial 130nm CMOS
Cell	8T high-V_T SRAM
Voltage	350-700mV
Leakage Power	12.29nW @ 320mV (Standby) 1.09nW @ 320mV (Shutdown)
Energy/access	6.24pJ/access @ 400mV
Special Features	1. High-V_T devices 2. Full-swing read 3. Read burst mode 4. RWL boosting to improve read stability 5. Read-before-write for half-select instability 6. WL boosting to improve write stability 7. Aggressive power gating for low power Standby & Shutdown modes

4.1.1 Bit-cell Array

Due to the challenges of operating the conventional 6T bit-cell at sub-threshold voltages, this array is made up of 64x128 8T bit-cells (Figure 2.3) with decoupled read and write ports. High-V_T devices are used within the bit-cells to reduce their leakage currents, and thus the standby power consumption of the array. However, since high-V_T devices have reduced on-current, the read and write margins are significantly degraded, necessitating the use of assist techniques to guarantee correct operation.

Read Operation

To read the 8T bit-cell, the read bitlines (RBL) are pre-charged by the column drivers before the read wordline (RWL) is asserted by the row driver. Depending on the data within the cell, RBL is either discharged or kept high. However, due to the reduced I_{ON}-to-I_{OFF} ratio in sub-threshold, the off-current in the unselected bit-cells on the same RBL might cause an incorrect value to be read out. Thus, the footer voltages (VVSS) of these unselected bit-cells are held high as in [12], and only the accessed bit-cell VVSS is discharged before a read operation. Since the VVSS signal is shared between bit-cells on the same row, its driver must

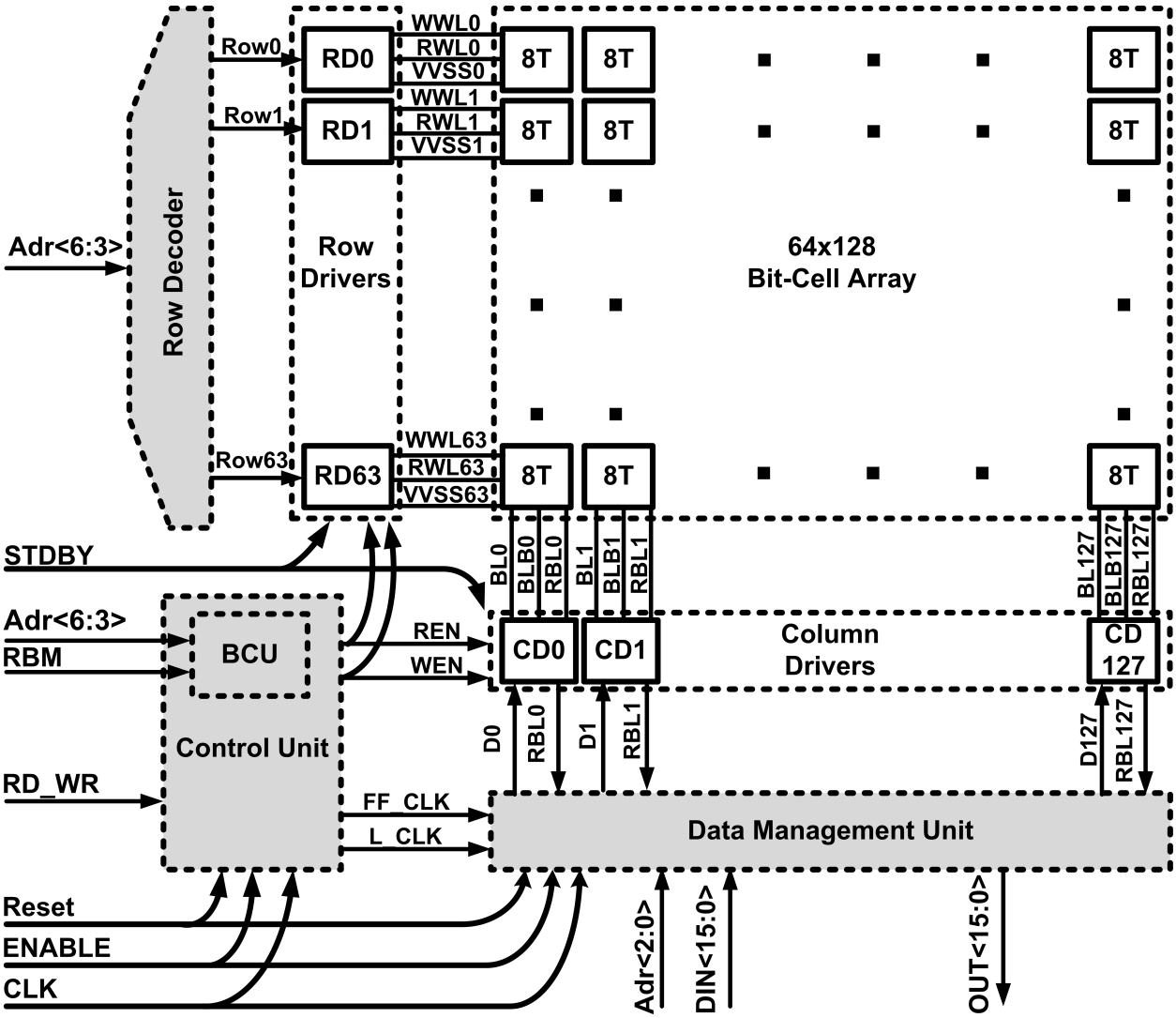


Figure 4.1: Array block diagram blocks in light gray are power gated during Standby mode.

be designed to sink the current from all the bit-cells in a row. Thus, the pull down network of the VVSS driver is overdriven by a charge pump circuit (area overhead $< 3\%$) to ensure the VVSS node does not rise [12].

Due to the reduced I_{ON} of the high-V_T devices, the read operation cannot be guaranteed across all process corners/temperatures without the use of an assist technique. Thus, two read assist techniques were considered to improve the read-ability of the 8T cell. In the first approach, RWL is boosted using the charge pump circuit introduced in [12]. The charge pump circuit was used instead of a level converter since it does not require an additional high supply

Table 4.2: Comparison between high-V_T read port with RWL boosting and nominal-V_T read port for read assist (based on chip measurements).

	Nominal-V_T read port	**High-V_T read port with RWL boosting**
Read frequency	114.7KHz @ 400mV	26.6KHz @ 400mV
Leakage Power per bit	3.4pW @ 320mV	1.5pW @ 320mV
Read Power per accessed bit	35.4nW @ 400mV	9nW @ 400mV
Area overhead (normalized)	1x	1.05x

voltage, and simulation results showed that it consumes less energy. In the second approach, nominal-V_T devices were used in the 2T read port instead of high-V_T devices and no boosting was performed. The two approaches were fabricated and their maximum frequency, leakage power and active power was compared. Table 4.2 shows the measurement results from the two arrays. Even though, the nominal-V_T read port improves the read frequency significantly, it also results in more than 2X increase in leakage power and ∼4X increase in active read power. Since this array targets IoT applications with relaxed frequency requirements, the RWL boosting approach was chosen instead of the nominal-V_T approach to ensure reduced leakage and active power.

To read the value on RBL, our implementation uses a simple output buffer instead of the commonly used sense amplifier to avoid the challenges of operating it at sub-threshold voltages. By limiting the number of bit-cells in a column to 64 and allowing RBL to completely discharge, the output buffer can correctly read the contents of the selected bit-cells. This increases the read energy but simplifies timing and increases robustness to variations.

To reduce the read power/energy, the array employs a read burst mode (RBM) feature which makes use of the fact that when RWL is asserted, the complete row experiences a read operation. Thus, when consecutive addresses in the same row should be read, it is enough to perform the read operation once, and save the data in latches for the consecutive reads. Accessing the latches will consume significantly lower energy than performing a normal read thus reducing the overall read energy. The Burst Control Unit (BCU) implementing RBM has negligible impact on the power (<0.7%), performance (0%) and area (<<1%) of the

system, and the potential savings it offers is significantly higher than the cost of implementing it. Section 4.1.2 describes the implementation of the RBM.

To further reduce the read power/energy of instruction memory ("read-mostly") arrays in IoT systems, SoC designers can make use of the fact that reading a "1" consumes significantly lower power than reading a "0" in 8T SRAM bit-cells [34]. The higher read "0" power is due to the discharging of RBL needed to read a "0"; whereas reading a "1" does not discharge RBL and thus the only contribution to the read power is the leakage current through unselected cells in the column which is kept at a minimum by boosting the unselected cell VVSS. By designing the instruction set of the IoT processor or by including an encoding scheme that results in more "1" bits than "0" bits in each word, the active power consumption of the array can be significantly reduced. The impact of these techniques on the system area, performance and standby power varies depending on the application. If the IoT processor instruction set is modified, the impact on area, performance and standby power is zero. However, this option might not be available to all system designers. On the other hand, the encoding scheme can be widely used but will have an impact on the area, performance and standby power of the system. For example, if the encoding scheme in [35] is used, the area overhead will be ~6% and the standby power will be increased by ~6%. This scheme will have little impact on performance but will allow 13% reduction in active power at 400mV, assuming it can reduce the number of "0" within a word from 50% to 25%. Table 4.3 below shows the difference in active power consumption of the array (not system) when different percentages of "0"s and "1"s are used within a word. Reducing the number of "0"s within a word from 50% to 25% results in 8% and 21% reduction in read power at 350mV and 400mV, respectively. Also, since our array relies on read-before-write to avoid half-select disturbs, the write power is reduced by 9.5% and 17% at 350mV and 400mV, respectively.

Table 4.3: Read and write power (in nW/KB) of the array containing different percentages of "0" and "1" bits in each word (based on chip measurements).

(nW)	0% "0" bits	25% "0" bits	50% "0" bits	75% "0" bits	100% "0" bits
Read Power @ 350mV	41.83	46.07	50.31	54.56	58.80
Read Power @ 400mV	83.07	113.48	143.89	174.3	204.71
Write Power @ 350mV	44.74	49.99	55.24	60.49	65.74
Write Power @ 400mV	123.89	156.09	188.29	220.49	252.69

Write Operation

When writing into the 8T bit-cell, the column drivers set the data on the write bit-lines (BL and BLB) before the row driver asserts the write word-line (WWL). Cells sharing the same WWL experience a half-select pseudo-read operation that might corrupt their contents. Thus, we adopted a row read-before-write (RBW) implementation, since it provided a good compromise between the added area needed to implement a different bit-cell topology, and the added power/energy and area needed to implement the additional logic and drivers for the one word per bank solution. Even though RBW will increase the energy consumed during a write operation, this increase is acceptable for "read-mostly" arrays where the number of writes is limited.

Since high-V_T devices are used, the write-ability of the cell is degraded due to the reduced drive strength of the pass transistors. Thus, a write assist technique is needed to guarantee correct write functionality. We evaluated the different write assist techniques to determine the optimal choice for our implementation. Since most battery-less SoCs do not require high performance, the static write margin (WM) can be used as an evaluation metric instead of the critical WL pulse width [31]. Column-based assist techniques such as NegBL and LCV_{DD} were not included in the evaluation due to the large area and energy overhead they will incur when the complete row is written in a row RBW implementation. On the other hand, row-based assist techniques such as WL Boosting and V_{SS} Raising can improve the margins to allow sub-threshold operation with limited impact on area and power. Figure 4.2 shows the impact of the row-based techniques on the 3σ WM of the 8T bit-cell at the SF corner (worst write corner) with temperature set to 25^0C. For the same degree of assist applied, WL

boosting shows more improvement in WM, and reduces the minimum voltage (write V_{MIN}) at which a write operation can be successfully completed. Thus, the WL boosting assist technique was adopted in this design. To implement this boosting, a charge pump circuit [12] was added within the row driver circuit (RDx) to boost WWL during a write operation.

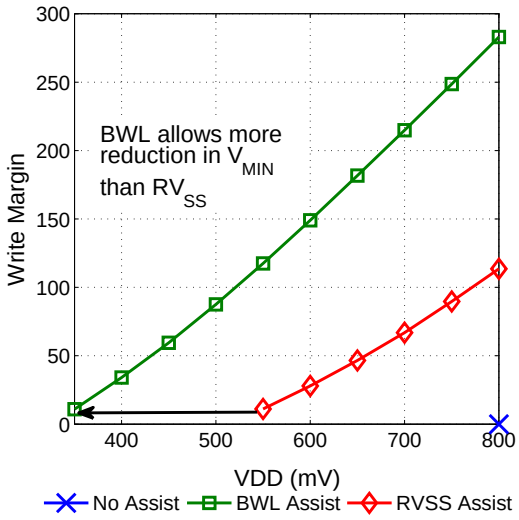


Figure 4.2: The 3σ Write Margin with WL Boosting (BWL) and V_{SS} Raising (RVSS).

4.1.2 Control and Data Management Units

The control unit is responsible for reading the inputs to the array, determining the correct mode of operation, and generating the appropriate read, write, and control signals. The control unit takes three inputs: active mode (ENABLE), read/write (RD_WR) and read burst mode enable (RBM), and generates four output signals to control the array: read enable (REN), write enable (WEN), latch clock (L_CLK), and output register clock (FF_CLK). In active mode (ENABLE=1), the control unit is ready to read/write data into the SRAM array. When RD_WR is asserted, data is read out of the array. First, the control unit asserts REN which then controls the row drivers that set RWL and VVSS of each row to the appropriate values. At the end of the read cycle, L_CLK is driven high to provide the latching edge of the latches. These latches are used to hold the read data to be used when read burst

mode is enabled (RBM=1) or when a write operation should follow the read in the RBW implementation. Finally, FF_CLK is asserted to provide the edge for the output registers. The read operation is always performed on the high clock phase and takes only half a clock cycle to complete. Thus, the data is available in the latches by the end of the high clock phase. If a write operation is requested (RD_WR=0), a read is first performed during the high clock phase but FF_CLK is not toggled. At the falling edge of the clock, the control unit asserts WEN which then controls the row and column drivers to set WWL and BL/BLB.

When RBM is enabled, the burst control unit (BCU) within the main controller will keep track of the addresses being accessed and the RD_WR signal. Once two consecutive addresses in the same row are read, the BURST signal goes high indicating to the control unit that the data is already available within the latches, thus REN and L_CLK are not toggled.

The Data Management Unit (DMU) shown in Figure 4.3 manages the data flow in the array. The DMU contains the read output buffer, the data latches, the output registers and the logic required to choose between the input data and the latch data for the write operation. The DMU takes as input the read bit-lines ($RBL < 127 : 0 >$), the input data ($DIN < 15 : 0 >$), the column address bits ($ADR < 2 : 0 >$), L_CLK and FF_CLK, and outputs the data read from the array ($OUT < 15 : 0 >$) and the data for the write drivers ($D < 127 : 0 >$).

Figure 4.4 shows the timing diagram of a read operation followed by a write operation assuming the array is in active mode. The RBLs are precharged during the low phase of the clock (CLK). If the RD_WR signal is high at the rising edge of CLK, REN is driven high to start the read operation. The latch clock signal - L_CLK - is held low until the end of the read operation (signaled by REN going low) where it is toggled high to enable the data latches to save. Based on the column address, one of the tristate buffers in the DMU is enabled and passes the data to the rising edge triggered output registers controlled by FF_CLK. FF_CLK is driven low at the start of the read operation (REN=1) and high at the end of the read operation (REN=0). When the RD_WR signal is low at the rising CLK edge,

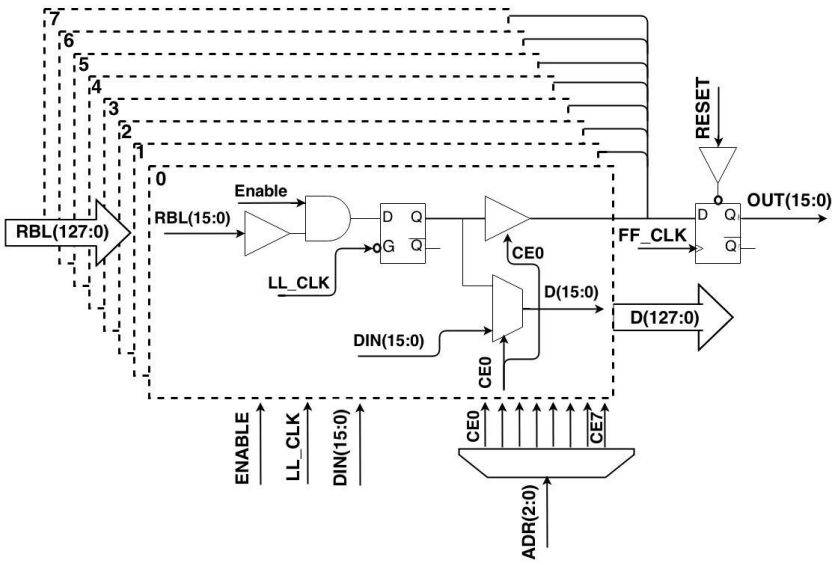


Figure 4.3: Data Management Unit (DMU) block diagram.

a write operation is performed. The write operation starts with a read on the high CLK phase (REN=1 and L_CLK=0). FF_CLK is not toggled since this data does not need to appear at the output. Once the row data is available (falling CLK edge), the multiplexers in the DMU choose between the latch data and the input data ($DIN < 15 : 0 >$) based on the column address, and then feed the result ($D < 127 : 0 >$) into the column drivers. Next, the WEN signal is asserted, to enable WWL and drive BL/BL to complete the write operation.

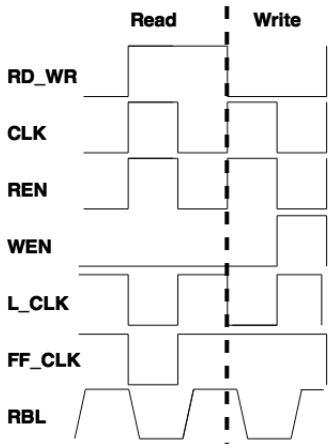


Figure 4.4: Timing diagram for read and write operations.

4.1.3 Power Reduction Features

To reduce the power consumption of the array, three low power modes were added. In the Hold mode, the SoC is not accessing the memory, thus the ENABLE signal is held low, and the clock signal to the memory is gated. When the SoC is in a low power state, the SRAM array can be either completely shut down or placed in a low power data retention (standby) mode. In the Shutdown Mode, the complete array is power gated, and the data is lost. In the Standby Mode, only the peripherals are power gated while the row and column drivers and the bit-cell array retain their state. The row and column drivers isolate the power gated circuits from the ON circuits when the STANDBY signal is enabled. The row driver will keep RWL and WWL held low and VVSS held high, and the column drivers will hold BL/BLB low and RBL high.

4.2 Chip Measurements

The proposed array was fabricated in a commercial 130nm bulk CMOS technology and tested at room temperature. The chip operates for both read and write broadly in the sub-threshold region between 350mV and 700mV (Figure 4.5), and can retain data down to 320mV.

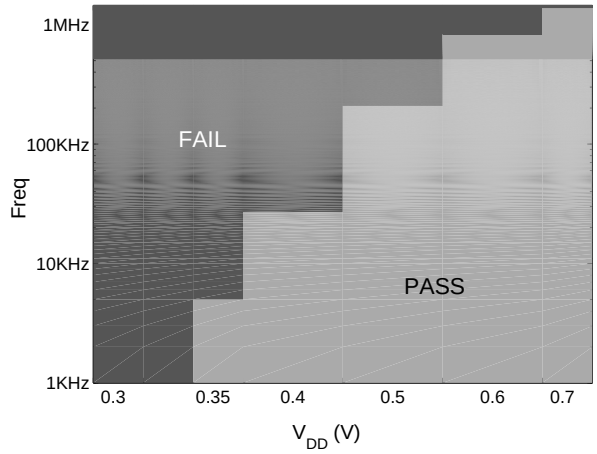


Figure 4.5: Measured shmoo plot of ULP SRAM.

Figure 4.6 shows the standby power consumption of the array in the three supported modes: Hold Mode, Standby Mode, and Shutdown Mode. The power consumption of the array is minimized at the data retention voltage of 320mV to 29.49nW, 12.29nW, and 1.09nW in the Hold, Standby, and Shutdown modes, respectively. The Standby and Shutdown Modes are particularly useful for instruction and data memories, respectively, in battery-less SoCs when energy harvesting resources are scarce.

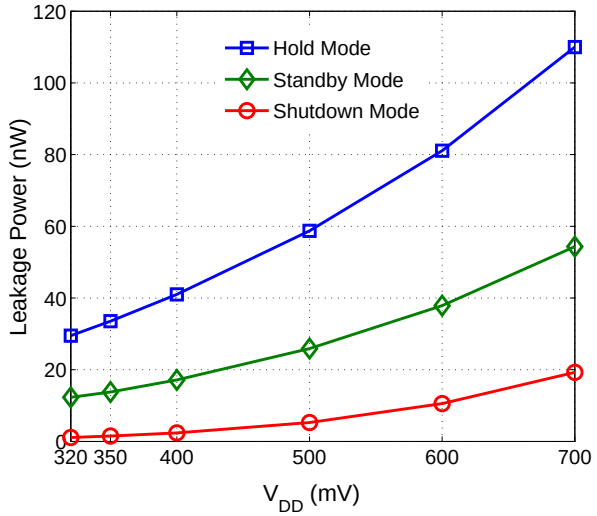


Figure 4.6: Measured power consumption during Hold, Standby and Shutdown modes.

The read and write energies (Figure 4.7) are minimized at 400mV to 5.41pJ/access and 7.08pJ/access, respectively. The read burst mode can provide up to 22% reduction in active read energy at 400mV when enabled. Measurement results show that the charge pump circuits used to boost RWL, WWL and VVSS consume only 3% of the total read/write power at 400mV.

Table 4.4 shows a comparison between the measured results of our chip and previously presented designs. The active energy per bit and leakage power per bit shown in the table take into account the energy/power consumed in the peripheral logic. The active energy per bit is calculated as the average of the read and write energies per bit. Our design shows one of the lowest standby power consumed per bit of memory at 1.5pW/bit with the lowest

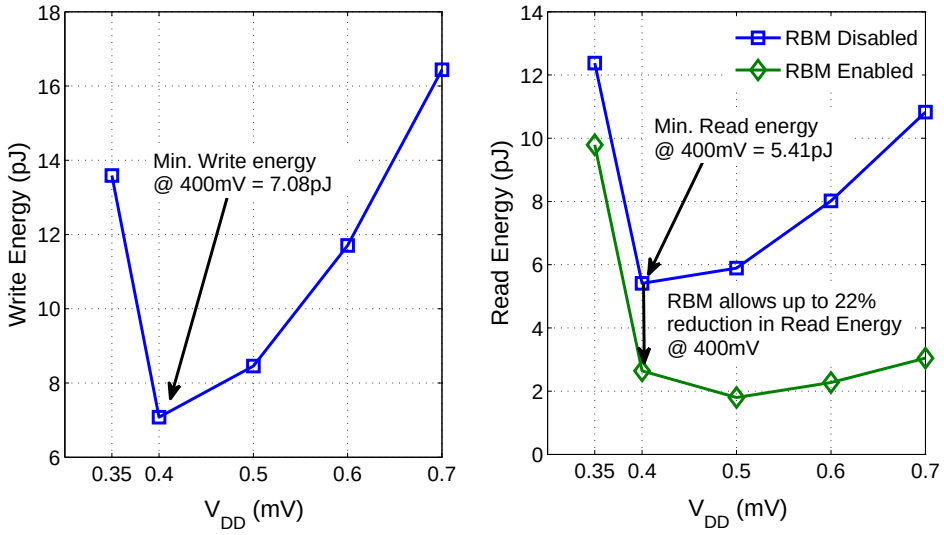


Figure 4.7: Measured write and read energy with read burst mode enabled and disabled.

number of transistors per bit-cell (8T compared to 10T [36] and 14T [14]), making it ideally suited for battery-less SoCs with multiple operating modes.

Table 4.4: Comparison with state-of-the-art arrays. *total energy reported in fJ since word size was not provided.

	Tech.	V_{DD}	Cell Type	Transistor Type	Array (Kb)	Word Size	Freq. (MHz)	Energy (fJ/bit)	Leakage Power (pW/bit)
This work	**130**	**0.32** **0.4**	**8T**	**High-V_T**	**8**	**16**	– **0.027**	– **390**	**1.5** **1.7**
[12]	65	0.35	8T	–	256	128	0.03	870	8.4
[13]	65	0.3 0.4	9T	Mixed V_T	2	32	0.22 2	18.2 21.8	17.8 25.4
[14]	65	0.5	14T	High-V_T	4	32	0.11	14	0.5 @ 0.22V
[36]	180	0.35	10T	Mixed V_T	24	32	0.053	–	0.0019
[37]	65	0.5	6T	Low Leakage	1024	–	250	–	5.7
[38]	65	0.4	8T	Low Power	64	128	~0.06	78	6.1 @ 0.25V
[39]	65	0.26	7T	–	32	–	1.8	5600*	–
[40]	90	0.23	Z8T	–	64	–	0.5	80000*	305

4.3 Individual Contribution

This work was performed as part of a team including Harsh Patel, James Boley and Arijit Banarjee. My contributions to the chip include:

1. designing and laying out the high-V_T cell.
2. simulating and verifying the extracted cell to determine read/write stability.
3. laying out the charge pump circuit used to implement the WL boosting assist.
4. helping with the design of the read burst mode.
5. laying out the complete 2KB array with its peripherals.
6. testing the high-V_T bank.
7. writing the paper on the high-V_T bank.

4.4 Conclusion

This chapter presented a 1KB SRAM chip fabricated in 130nm CMOS that operates between 350mV and 700mV for ULP sub-threshold operation. High-V_T devices are used within the 8T bit-cell in the array. Read and write assist techniques are introduced to guarantee correct operation. A read-before-write approach is implemented to address half-select instability. The read and write energy is minimized at 400mV. A read burst mode is implemented to reduce read energy when consecutive addresses are accessed and saves 22% active read energy. Increasing the percentage of "1" bits within a word allows significant reduction in both the read and write power. Aggressive power gating reduces the power consumption down to 12.29nW with retention and 1.09nW when data is not needed (at the data retention voltage - 320mV). Compared to the state-of-the-art ULP SRAMs, the proposed design gives the lowest full array leakage power per bit at 1.5pW/bit for an 8T bit-cell array.

Chapter 5

Low-V_T STT-RAM

[1] After reducing the power and energy consumption of the on-chip volatile memories, in this chapter, we look into non-volatile memories starting with STT-RAM. STT-RAM has the potential of enabling embedded non-volatile memory with ∼2-4X smaller cell size than SRAMs. The non-volatile advantage of STT-RAM implies near zero array leakage power during standby, which makes them ideal for low power circuits. In this chapter, we present a methodology to reduce the write energy of STT-RAM using a combination of cell-level and array-level techniques, and at the same time improve the write-ability of the cell.

5.1 Low-V_T Cell

As we have previously described in Section 2.3, STT-RAM cells require a large current during a write operation. Thus, to improve the write-ability of the STT-RAM, we propose lowering the V_T of the NMOS access transistor (Figure 2.4a) to allow larger current to pass through, while avoiding drawbacks of state-of-the-art schemes in terms of violating NMOS gate-oxide reliability and increasing the cell area. The increased current can be traded for (i) smaller cell area for the same write speed, (ii) faster write-time for the same cell area, (iii) lower write V_{MIN} (energy) for the same speed and area and/or (iv) larger MTJ area and thus higher

[1] This chapter is based on [FBY3]

thermal stability for the same cell area (assuming the cell area is not determined by the MTJ) and write speed. A low-V_T access transistor can be used along with other write assist techniques such as WL boosting if needed, to further improve the write-ability of the cell.

Before discussing the proposed low-V_T cell, we introduce the write-margin as a metric to quantify the impact of using low-V_T devices on the write operation. The write-margin is defined as the difference between the MTJ write current and the critical switching current (either I_{C+} or I_{C-}). To verify low-V_T write margin and area advantage, simulations of a write-"1" operation were performed using a commercial 32nm process at $V_{DD} = 1.05V$ and typical NMOS corner with temperature set to 110^0C while taking into account 6σ local variations in both the NMOS and the MTJ. The width of the NMOS and its V_T were varied, and the MTJ current was measured and compared to the critical current (I_{C-}). Figure 5.1 plots the worst case write-"1" margin as a function of NMOS width and V_T. Reducing V_T by 200mV gives 18X improvement in the write-margin for the same NMOS width. Alternatively, for the same write margin, the cell with 200mV lower V_T occupies 1.8X less area than the nominal V_T cell (assuming cell area tracks access transistor width to a first order).

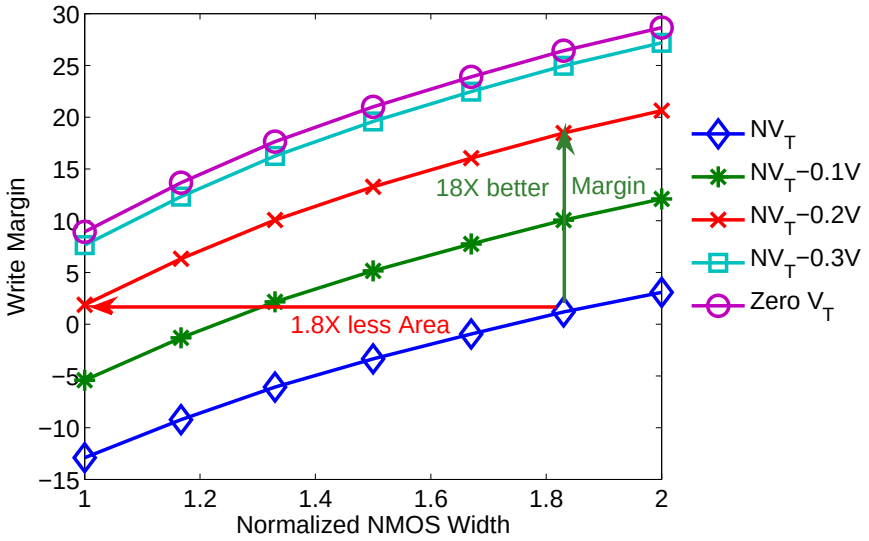


Figure 5.1: Impact of V_T and width on the write-margin.

5.2 All-Digital Programmable Driver

One consequence of using a low-V_T NMOS is its higher leakage current which will increase the array read/write energy. This extra leakage current can introduce a number of challenges in unselected bit-cell on the same column as the selected bit cell as illustrated in Figure 5.2. For a given data polarity and with process variations, the high leakage current through the unselected cell can accidentally switch its state, resulting in a false write (FW) scenario. FW are much more likely to occur during a write-"0" operation, since the required critical current (I_{C+}) is much lower and the NMOS is in a source follower configuration. Furthermore, for a selected cell, which is not write-limited (not experiencing a P-to-AP write operation), the large current passing through it can violate the oxide breakdown (V_{BD}) limit of its MTJ. V_{BD} violations are more likely to occur when the cell is holding a "0" and a write-"0" operation is taking place.

To address the three challenges (FW, V_{BD} violation and increased power) while maintaining the write reliability of a low-V_T STT-RAM, the SL/BL voltages can be raised during a write-"0"/write-"1" operation. Since the write-"0" operation requires less critical current than the write-"1" operation ($I_{C-} > I_{C+}$), the current through the cell can be reduced by inducing a source follower configuration. This can be achieved by raising the SL voltage until the on current supplied to the worst selected cell is just enough to write the cell (write-"0" margin is close to 0). Besides reducing active write energy, a considerable reduction in the leakage current in unselected cells is attained. In addition, raising SL ensures that V_{MTJ} of the selected cell remains well below the V_{BD} limit. Figure 5.3 shows the on and leakage currents when nominal SL and raised SL are used for different V_T shifts from nominal ($\Delta V_T = 0$ refers to nominal V_T case).

To raise the voltage on BL/SL, the new all digital and programmable write driver in Figure 5.4 is proposed. The strength of the pull-down network of the write-driver is adjusted to create different SL (BL) bias voltages. This is done by changing the pull-down stack height, as well as combining NMOS pull-downs with different device types such as PMOS pull-down,

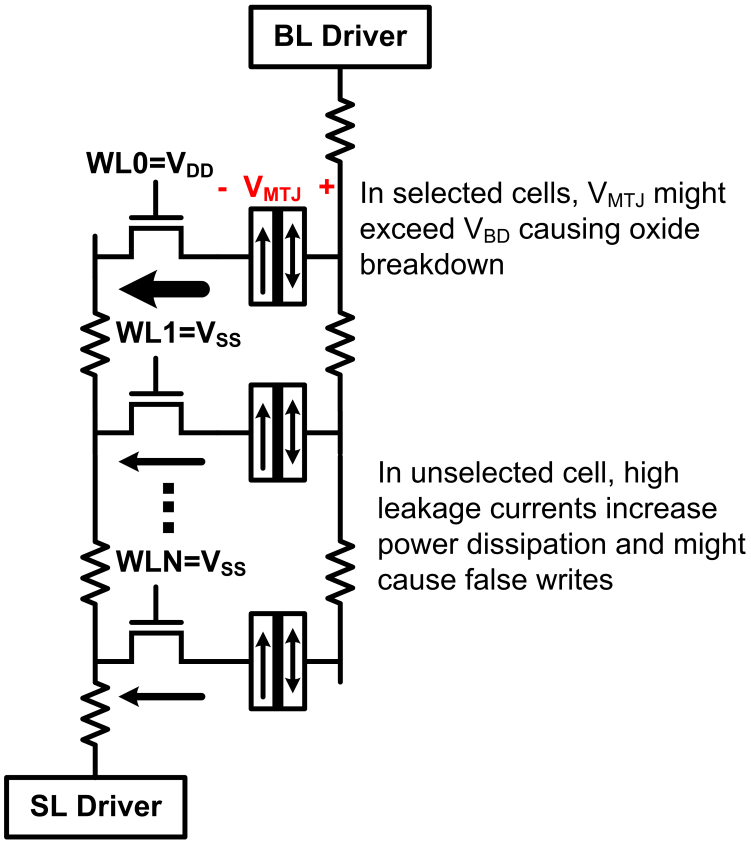


Figure 5.2: Large leakage current through unselected cells and high V_{MTJ} in a selected cell during write-"0" case.

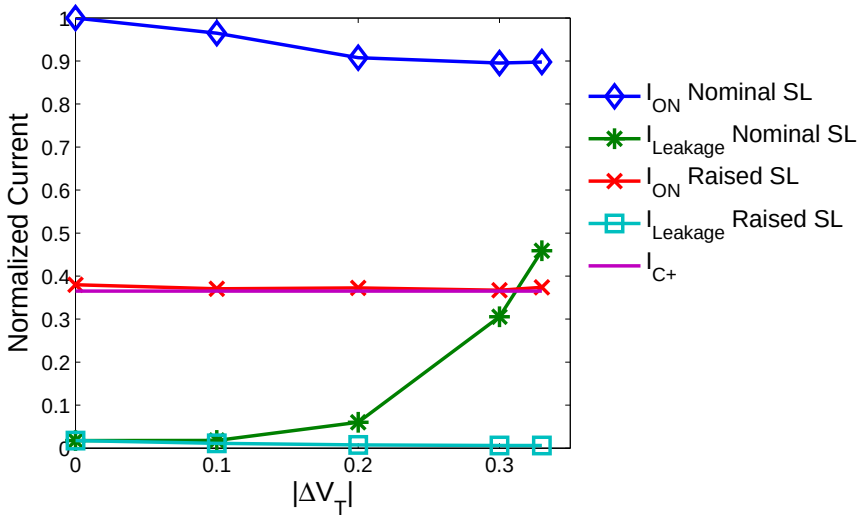


Figure 5.3: ON current and leakage current during a Write-"0" operation with and without raised SL.

a high-V_T NMOS, and/or a diode-connected NMOS. Note that transistors within a given stack may have different sizes than those in other stacks. A given stack is selected (using signals P0, P1, P2) per die depending on the die process corner and under the worst case temperature.

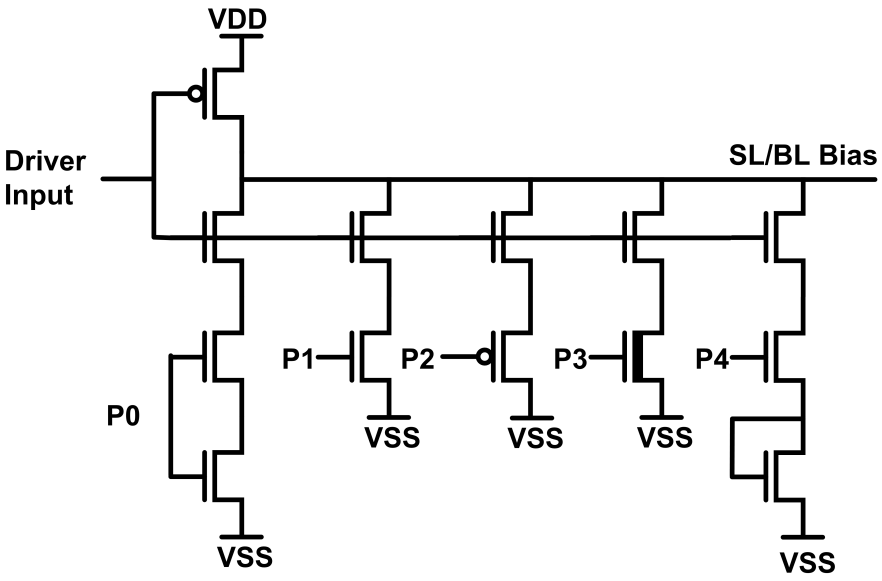


Figure 5.4: Proposed write driver for SL/BL bias generation.

5.3 Design Methodology

In order to facilitate the design of the proposed cell and driver, a design methodology is presented in Figure 5.5. The main aim of this procedure is to minimize the write energy while ensuring correct write-"0"/"1", avoiding FW and V_{BD} violations and minimizing the cell area.

In the first step, the minimum NMOS width that guarantees correct write-"1" operation across all process corners (T, F, S), temperatures (0^0C, 110^0C), and under the worst case condition (taking into account $6\sigma+$ local variations) is determined. As shown in Table 5.1, using a low-V_T device allows significant reduction in the cell area (up to 60% reduction).

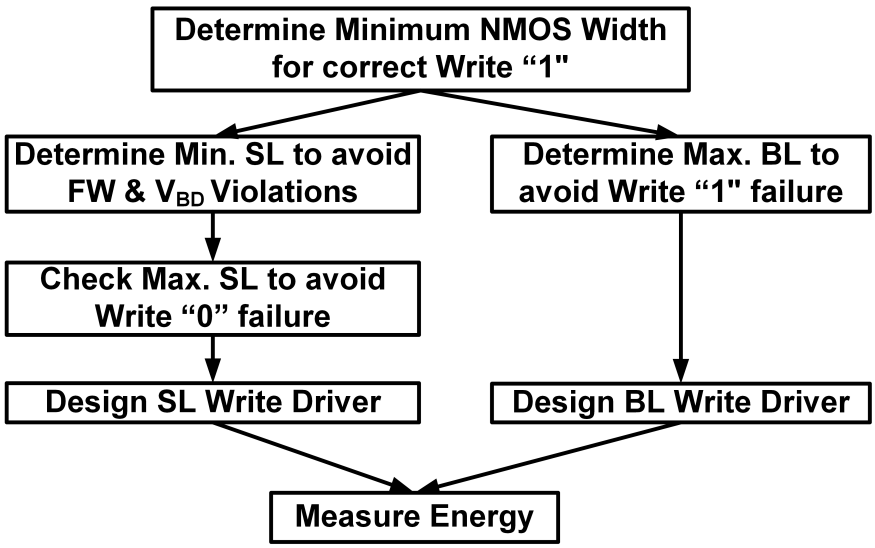


Figure 5.5: Design methodology for low V_T STT-RAM with programmable driver.

Table 5.1: Normalized NMOS width for correct write "1".

ΔV_T	**0 (Conv.)**	**-0.1**	**-0.2**	**-0.3**
Normalized Width	1	0.67	0.47	0.4

Once the NMOS size is chosen, the strength of the pull-down path of the BL and SL drivers must be determined carefully to ensure that the write operation remains successful while the adverse effects are addressed. Thus, the SL voltage is first varied to determine the minimum bias that would eliminate FW and V_{BD} violations under the worst case conditions. First, we determine the conditions that make FW most likely to occur. A write-'0" operation forces the NMOS transistors in a column into a common source configuration which increases their leakage current. Thus, an unselected cell holding a "1" is most likely to experience FW since a smaller current is needed to switch its content. The physical location of this cell will also impact the current passing through it due to interconnect parasitics. The cell physically closest to the SL Driver (Cell N, Figure 5.2) sees the least parasitics on SL and thus will have the highest leakage current and chance of experiencing a FW. Also, in the worst case, all other cells in the column hold a "1" which will steer most of the leakage current into the worst case cell. After determining the worst case cell for FW, the minimum required SL

voltage required to eliminate FW for each corner and temperature while taking into account $6\sigma+$ local variations is determined as FW SLmin.

Next, the worst case cell for V_{BD} violation is determined. A write-"0" operation into a cell already holding a "0" will induce a higher current through its selected MTJ inducing a high voltage across it. The cell physically closest to the SL driver sees the least parasitics on SL and thus the least parasitic-induced SF. Also, in the worst case, all other cells in the column hold a "1" which will steer most of the current into the worst case cell aggravating the problem. The minimum SL bias that addresses V_{BD} violation for this cell is determined for each corner and temperature while taking into account $6\sigma+$ local variations as V_{BD} SLmin. The lower bound on SL is then determined by taking the maximum of the FW SLmin and the V_{BD} SLmin.

After finding the lower bound on SL, the upper bound is determined by considering the write-"0" operation into the worst case cell for each process corner and temperature while taking into account $6\sigma+$ local variations. In this test, the worst case cell is the one that sees the most interconnect parasitics (Cell 0, Figure 5.2) during a write-"0". Under worst case conditions, all unselected cells hold "0" to steer as much of the driver current away from the accessed cell as possible.

Following this procedure, the two bounds on SL for different V_T values are shown in Table 5.2. It is worth noting that FW is dominant for low-V_T cells where the leakage current is high; whereas V_{BD} violations is more dominant with higher V_T cells. While this might seem counter intuitive, the extra leakage through the unselected cells with low-V_T devices reduces the on current passing through the selected cell thus reducing V_{MTJ} below V_{BD}.

After determining the bounds on the SL bias, a simpler procedure could be used to determine the bounds on the BL bias during a write-"1" operation. FW and V_{BD} violations are less likely to occur during a write-"1" operation due to the source follower configuration of the NMOS. Thus, the only limit on BL is imposed by a successful write operation. During a write-"1", Cell N (Figure 5.2) will see the most interconnect parasitics, and thus the highest

Table 5.2: Allowable range of SL bias required to ensure no V_{BD} violations (black) and no FW (gray).

ΔV_T	0(Conv.)	-0.1	-0.2	-0.3
T-110^{0}C	0.05-0.3	0.05-0.33	0-0.34	0-0.39
T-0^{0}C	0.15-0.29	0.05-0.32	0-0.34	0-0.38
F-110^{0}C	0.10-0.37	0.05-0.39	0-0.39	0.1-0.46
F-0^{0}C	0.15-0.36	0.1-0.38	0.05-0.39	0.1-0.46
S-110^{0}C	0.05-0.26	0-0.28	0-0.28	0-0.32
S-0^{0}C	0.10-0.25	0.05-0.27	0-0.28	0-0.32

BL bias making it the worst cell to write-"1" into. Unselected cells on the same column holding "0" will steer current away from the cell being written contributing to the worst case condition. Once the limits on the BL/SL voltages are determined, the pull-down stack of the proposed driver can be designed for each of the process corners. To illustrate the advantages of the proposed driver, nominal V_T NMOS devices of different widths were used in the pull-down of the SL driver.

The final step is to measure the write energy. The average energy consumed by a column of STT-RAM cells including its write driver is measured for the different possible write operations (writing "0" or "1" into a cell holding "0" or "1"). Two cells are considered while measuring the energy: the top cell (Cell 0) and the bottom cell (Cell N) and the average of the two measurements is reported. Then, the probability of each write operation occurring is calculated and used to determine the average write energy according to the equation below:

$$\begin{aligned} E_{Total} = {} & E_{1-to-0} \times Pr(Writing\ 0\ \&\ Cell\ holds\ 1) + \\ & E_{0-to-1} \times Pr(Writing\ 1\ \&\ Cell\ holds\ 0) + \\ & E_{1-to-1} \times Pr(Writing\ 1\ \&\ Cell\ holds\ 1) + \\ & E_{0-to-0} \times Pr(Writing\ 0\ \&\ Cell\ holds\ 0) \end{aligned}$$

The above equation could be further simplified assuming that having a cell holding a "0" is equiprobable to having a cell holding a "1" and writing a "1" is equiprobable to writing a

"0".

$$E_{Total} = (E_{1-to-0} + E_{0-to-1} + E_{1-to-1} + E_{0-to-0})/4$$

As shown in Table 5.3, using the proposed driver to generate the BL and SL voltages during write-"1"/write-"0" operations reduces the overall write energy of the conventional (nominal V_T) cell by an average of ∼35% over process and temperature corners and by 37% for the low-V_T cell with $\Delta V_T = -0.2$. Most of this reduction is coming from write-"0" energy reduction (∼32% for the conventional cell and 30% for the low-V_T cell with $\Delta V_T = -0.2$) since the cell is originally designed (in terms of NMOS sizing and V_T selection) to be write-"1" limited.

Table 5.3: % cell write energy reduction when using the proposed driver as compared to a conventional design.

ΔV_T	0(Conv.)	-0.2	-0.3
T-110^{0}C	-34.94%	-41.68%	-33.16%
T-0^{0}C	-39.18%	-44.14%	-36.44%
F-110^{0}C	-37.52%	-39.35%	-39.00%
F-0^{0}C	-41.62%	-43.63%	-41.72%
S-110^{0}C	-25.61%	-24.02%	-16.94%
S-0^{0}C	-28.24%	-27.42%	-24.35%
Average	**-34.87%**	**-37.12%**	**-32.49%**

Finally, it is worth noting that the low-V_T cell might suffer from a degradation in read margins. To study the impact of low-V_T cells on the read operation, a basic current mirror sense amplifier is employed. The sizing of the amplifier was chosen to ensure the same read disturb margin for the conventional and the low-V_T cells. During read operation, all the pull-down paths in the SL driver are turned on to keep the SL bias as low as possible. For the used sense amplifier, the conventional cell shows 12% better distinguishability margin than low-V_T cell with $\Delta V_T = -0.2$; due to the extra leakage from the unselected cells. To improve the read margin, an optimized sense amplifier with offset compensation must be used.

5.4 Conclusion

In this chapter, we presented a hybrid array-level (programmable write-driver) and cell-level (low-V_T access transistor) technique to improve the write-ability of the STT-RAM cell. The proposed low-V_T cell provides higher write currents for the same size or lower cell area (1.8X) for the same write current. A methodology to design the proposed write-driver was introduced where the bounds on the BL/SL voltage during write-"1"/"0" are determined, then used to tune the pull-down stack of the driver to track the maximum BL/SL bias that reduces the write energy (by up to 37%) and ensure no false writes and MTJ breakdowns occur.

Chapter 6

Ferroelectric Auto-Recovery (FeAR) Sub-system

While the techniques in the previous chapter improve the write-ability of STT-RAM without increasing its energy consumption, these techniques result in slightly degraded read-ability. In this chapter, we will look into another type of non-volatile memory - Ferroelectric RAM (Fe-RAM). Unlike STT-RAM, we had the opportunity to fabricate a chip with Fe-RAM technology. Thus, we developed a non-volatile backup sub-system to complement the battery-less SoC. This chapter will introduce the ultra-low power FeAR sub-system consisting of a ferroelectric non-volatile memory with a specialized ultra-low power bus (ULP-BUS) interface to the SoCs. The proposed memory holds program and critical system data during power outages, and the ULP-BUS is designed to allow the integration of FeAR and the SoC in a compact System-In-Package (SiP).

6.1 System Overview

As was shown in Section 1.1, the amount of harvested energy varies significantly with environmental conditions and type of harvester. Thus, battery-less SoCs must adapt to variations in their power budget and recover after a complete power loss. While non-volatile

memories such as Fe-RAM and STT-RAM can retain their data and enable complete recovery, they consume significantly higher current during read/write than the volatile SRAMs making their use as instruction memories in battery-less SoCs impractical. Thus, to keep the system within the power budget, we propose adding a complementary non-volatile sub-system that can continuously hold the SoC program but is only read in the event of complete power loss. In this manner, the extra power consumed by the non-volatile memory is amortized over the frequency of power loss.

To enable recovery, the FeAR sub-system contains two non-volatile Fe-RAM memory arrays. The first holds the programming data of the SoC, while the second holds critical information the SoC wants to recover. A four-pin programming interface allows users to program the instruction memory on FeAR. An ultra-low power parallel bus (ULP-BUS) enables data exchange between FeAR and the SoC. On the SoC side, a power monitor (PM) and a cold-boot management system (CBMS) are implemented to facilitate back-up and recovery. The PM keeps track of the available energy and notifies the main controller when the energy drops below the critical level. The main controller then collects all the critical data and sends it to the CBMS which initiates back-up into FeAR.

Upon a power-on reset, the PM tracks the available energy and keeps the SoC in standby mode and FeAR in off-mode until the available energy exceeds the bootup threshold. Once enough energy is available, the PM instructs the CBMS to recover the programming and critical data from FeAR. After a bootup command is issued to FeAR, CBMS keeps track of the incoming data and programs the SRAM memory.

6.2 System Architecture

Figure 6.1 shows a block diagram of the proposed FeAR subsystem and how it connects to the ULP SoC. FeAR acts as a slave device to the SoC, relying on the SoC for power (VDDH and VDDL) and main control signals. The memory sub-system within FeAR consists of a 128x128

array and a 16x8 FIFO. The 128x128 array serves as a ferroelectric programmable read-only memory (Fe-PROM) to the SoC, while the ferroelectric FIFO (Fe-FIFO) is a read/write memory for the SoC to store any critical data. The Fe-PROM is programmed through a four-pin serial interface and read by the SoC through the ULP-BUS upon power up. In a standard system, the Fe-PROM is programmed before deployment, however, it can be re-programmed at any time using the four-pin programming interface. On the other hand, the Fe-FIFO cannot be accessed directly on the bench but can be written and read by the SoC through the ULP-BUS. The Fe-PROM, Fe-FIFO and their peripheral circuitry (write drivers, sense amplifiers and address decoders) make up the high power domain of FeAR running on VDDH. Meanwhile, the Control Unit (CU) and the ULP-BUS make up the low power domain running on VDDL (typically in sub-threshold). This arrangement was chosen to ensure reliable read/write operations to the Fe-RAM arrays while maintaining low power operation. This architecture also reduces the power consumed by level conversion through minimizing the number of signals requiring it.

6.2.1 ULP-BUS: Interface to the SoC

The ULP-BUS[1] is made up of a serial clock (SCLK), a master-out slave-in (MOSI), and an 8-bit wide slave-out master-in bus (MISO-BUS). The size of the ULP-BUS was chosen as a compromise between integration into an SiP and communication time and power. Each transmitter contains a digital CMOS driver which can be configured with a four bit binary input to control the drive strength of the pad. Each receiver contains a low swing sense amplifier. For minimum energy transmission, the driver is configured to create a partial swing output between VSS and the minimum detectable offset within a single clock cycle.

Figure 6.2 shows a block diagram of the interface between the ULP-BUS and the CU. The CU manages read and write into the memories based on inputs from the ULP-BUS. The ULP-BUS receives 8-bit serial commands from the SoC through SCLK and MOSI and

[1]The circuits in this block were designed by Christopher Lukas.

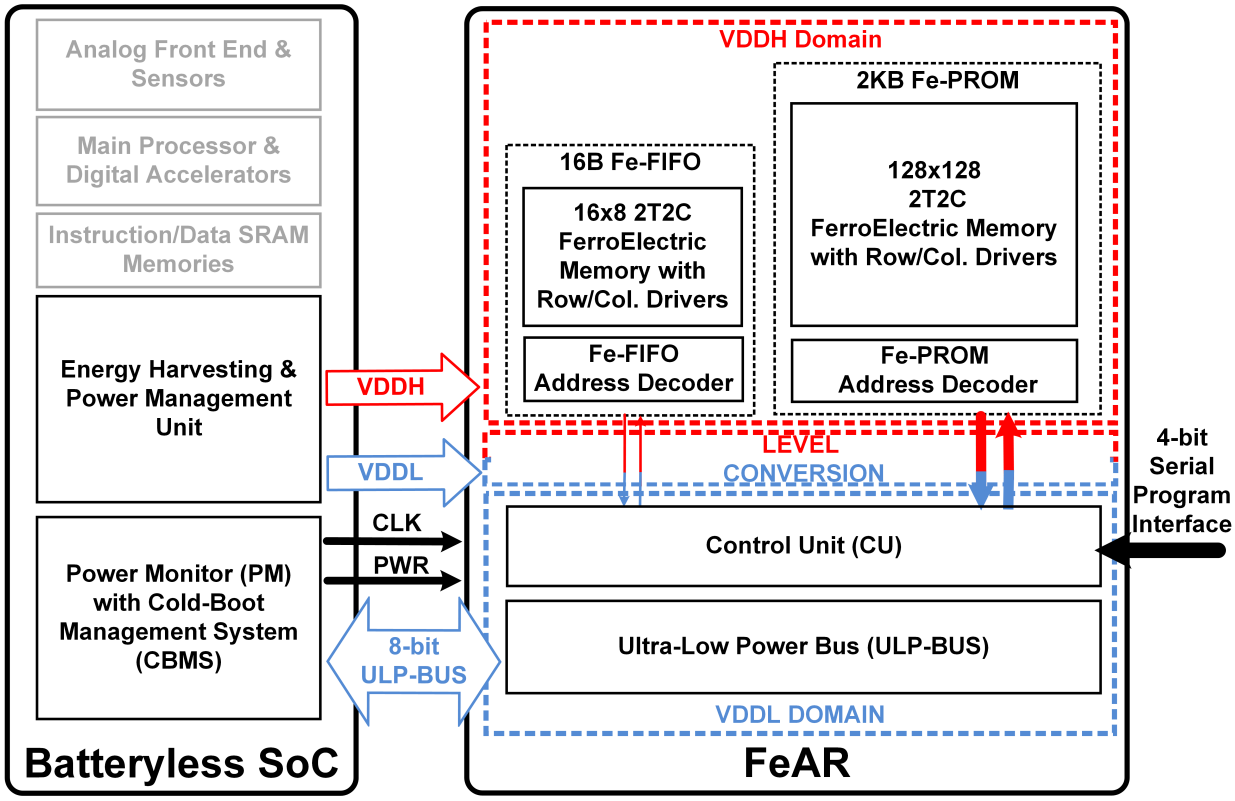


Figure 6.1: Block diagram of the proposed FeAR sub-system and its interface to a battery-less SoC.

transmits data back to the SoC through MOSI-BUS. Even though the SoC provides both SCLK and CLK, the two signals might be out-of-phase, requiring synchronization between the ULP-BUS and the rest of FeAR. Simple S-R latches are used to pass command signals from the ULP-BUS to the CU and status signals from the CU back to the ULP-BUS. The ULP-BUS allows three main commands: BOOTUP, BACKUP, and TEST. The BOOTUP command sends the contents of the Fe-PROM and the Fe-FIFO to the SoC and is usually received upon a power-on-reset. The BACKUP command saves data to the Fe-FIFO from the SoC and is usually received when the available power drops and critical data needs to be backed up. The TEST command is a debug feature that allows the testing of each memory and the ULP-BUS.

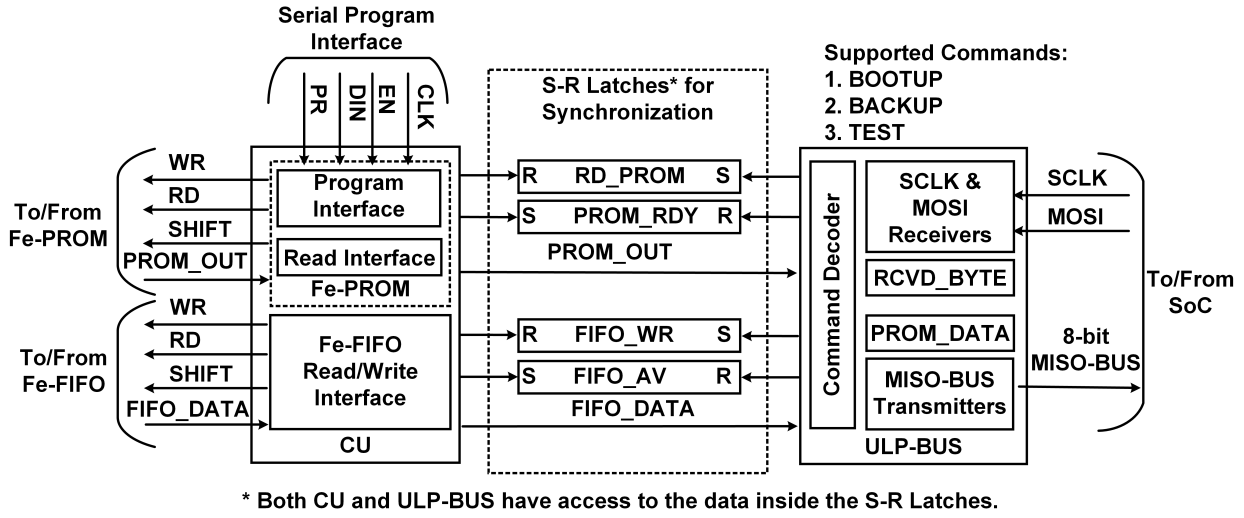


Figure 6.2: Block diagram showing the interface between the Control Unit (CU) and ULP-BUS

6.2.2 Array Architecture

At the heart of FeAR is a two transistor two ferroelectric capacitor (2T2C) Fe-RAM cell designed to hold the data and its complement. The 2T2C cell is chosen to improve the reliability by eliminating the need for a referenced sense amplifier. The differential nature of the cell will also allow low voltage read operation which reduces the power consumption of the complete sub-system. A simple reference-less sense amplifier is implemented to read out the cell contents without the need for a special enable signal, thus eliminating the need for complex control circuitry. The combination of the 2T2C cell and the sense amplifier enables read operation at voltages as low as 700mV (based on simulation results), resulting in considerable power savings. Figure 6.3 shows the structure of the Fe-PROM, the 2T2C cell, and the sense amplifier.

The row size (R) of the Fe-PROM array was chosen to reduce its read energy. A 2KB Fe-PROM array is implemented to match the program memory of the battery-less SoC. During a read operation to the Fe-PROM, one row is accessed and saved into a temporary register for the ULP-BUS to transmit to the SoC. The Fe-PROM array is then held in standby until all bytes are transmitted to the SoC (R/8 cycles). With this read methodology, different array structures are simulated to identify the one consuming the least average array read

power with a peak power within what the SoC's power manager can provide (∼1mW). In this simulation, one row of the array is read during the first cycle and held in standby for (R/8-2) cycles. Then, the average power over these cycles is calculated. The 128x128 array structure showed a good compromise between average power and peak power. Thus, the Fe-PROM memory was constructed as a 128x128 array.

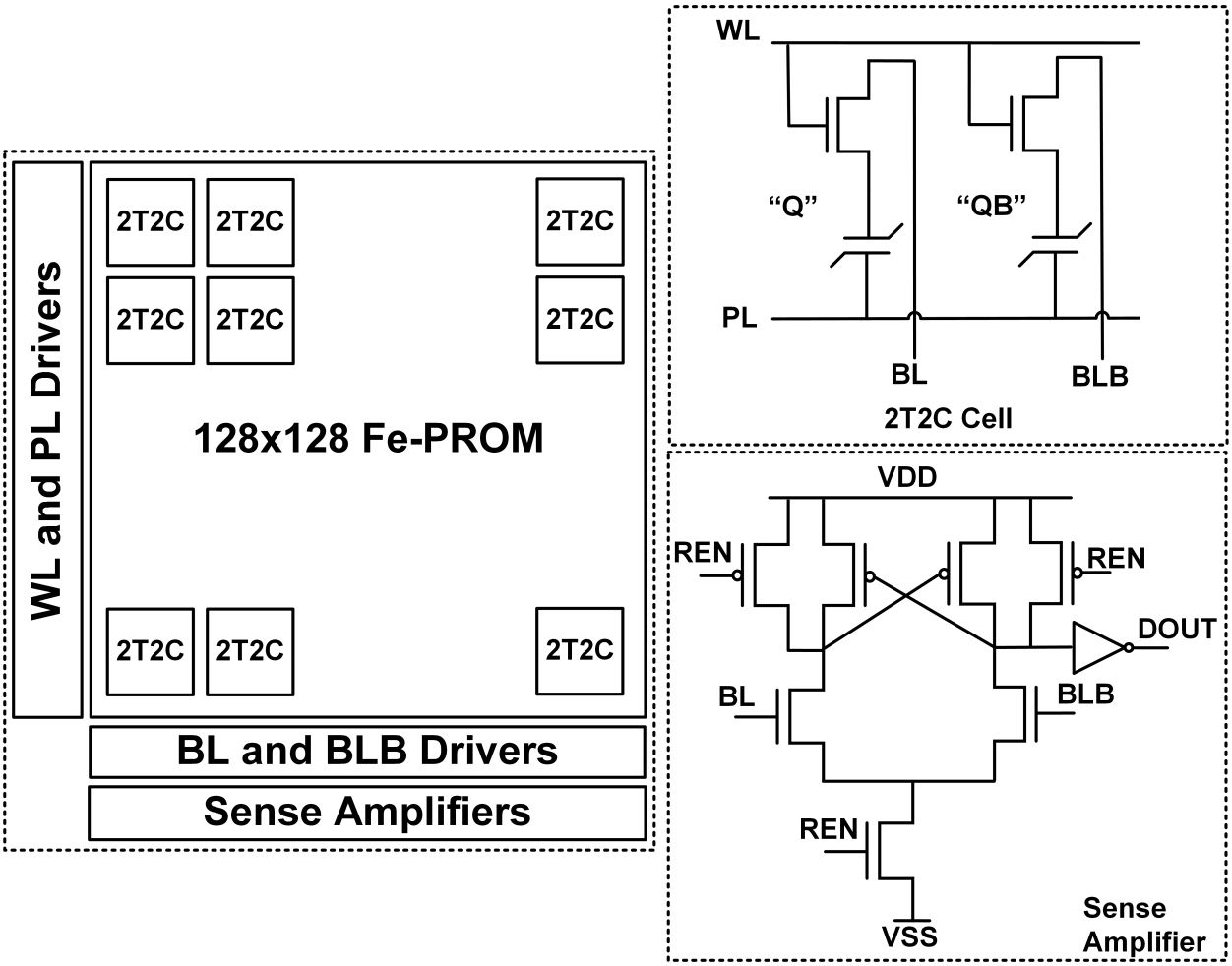


Figure 6.3: Array structure within FeAR showing the 2T2C cell and the reference-less sense amplifier.

Since these arrays are only used as backup for the SoC, they do not need to be randomly accessed. Thus, the address decoders of the Fe-PROM and Fe-FIFO are designed as rotating shift registers initialized to allow access to the first row. The rotating feature simplifies the test procedure. To further reduce the power consumed by the system, only the memory

arrays and their corresponding peripheral circuits are supplied from the high voltage domain (VDDH). Level shifters are inserted on the interface between the arrays and the control unit (CU) to facilitate information transfer between the two power domains.

6.2.3 FeAR Programming and SoC Recovery

Figure 6.4 shows the programming sequence for the Fe-PROM. To begin programming, the enable pin (EN) is asserted and the row data is shifted one bit at a time at the data pin (DIN). Once the complete row data is shifted in, the program pin (PR) is asserted for two cycles. During those two cycles, a row of the Fe-PROM is written (WR=1), and the address decoder shift register is rotated once (SHIFT=1) to enable access to the next row. To signal the end of the program, an end-of-file (EOF) word is written into the memory.

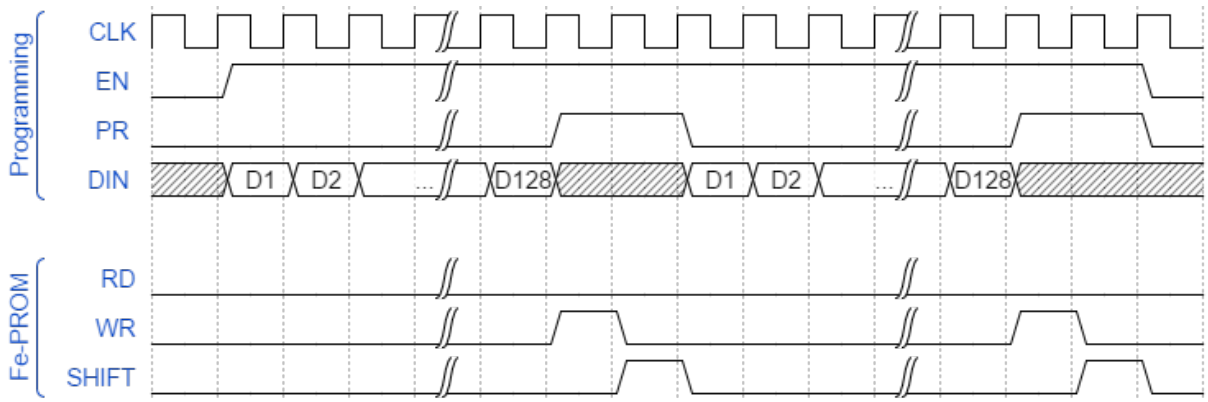
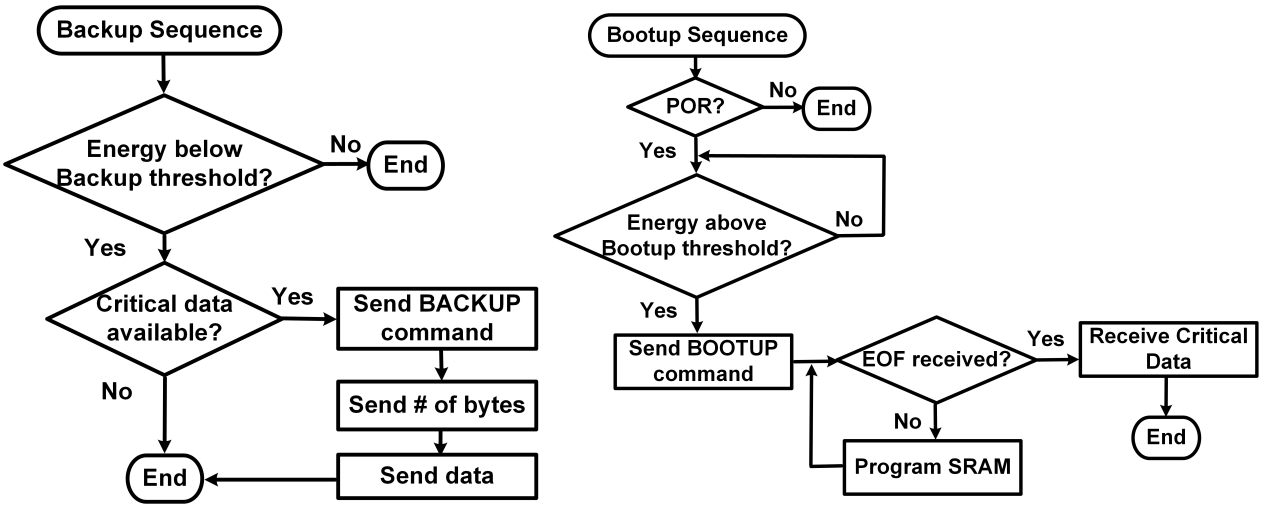


Figure 6.4: Fe-PROM programming waveform.

While the Fe-PROM can be programmed externally through the four-pin interface, the Fe-FIFO can only be accessed through the SoC. Figure 6.5a shows the backup sequence of the SoC. When the power monitor (PM) on the SoC detects a droop in the available power, it instructs the SoC to perform a back-up of the critical data into the Fe-FIFO. Once this data is ready, the SoC sends a BACKUP command through the ULP-BUS to FeAR followed by the number of bytes to be expected. The ULP-BUS then reads MOSI into RCVD_BYTE and sets FIFO_WR after receiving a byte of data (Figure 6.2). The CU then writes the contents of RCVD_BYTE into the Fe-FIFO and keeps track of the size of Fe-FIFO.

Figure 6.5b shows the bootup sequence of the SoC. Upon a power-on reset, the ULP-BUS receives a BOOTUP command from the SoC. It then prepares FeAR to send the programming and critical data to the SoC. Within FeAR, the ULP-BUS sets the RD_PROM signal to start reading the Fe-PROM (Figure 6.2). Once the CU detects a high signal on RD_PROM, it reads a row of the Fe-PROM into PROM_OUT and asserts the PROM_RDY signal. At the same time, it checks the number of bytes in the Fe-FIFO. If that number is non-zero, the CU reads the Fe-FIFO one byte at a time, saves the contents into FIFO_DATA, and asserts FIFO_AV. When the ULP-BUS detects a high signal on PROM_RDY, it saves PROM_OUT into PROM_DATA and de-asserts PROM_RDY. After resetting PROM_RDY, the ULP-BUS transmits PROM_DATA to the SoC 8 bits at a time starting from the least significant 8-bits. In the meantime, CU counts 16 cycles, reads another row of the Fe-PROM into PROM_DATA, and sets PROM_RDY. This procedure repeats until the CU and the ULP-BUS detect the end-of-file (EOF) word in the Fe-PROM. Once EOF is reached, the ULP-BUS sends the number of bytes in the Fe-FIFO and checks FIFO_AV. If it is set, the ULP-BUS sends the contents of the FIFO_DATA one byte at a time to the SoC.



(a) Backup sequence.

(b) Bootup sequence.

Figure 6.5: Backup and bootup sequences on the SoC.

6.3 Chip Results

FeAR was developed in a commercial 130nm ferroelectric technology. Two revisions of this chip were taped out. In the first revision, a bug in the pad ring prevented us from testing the system. In the second revision, bugs were fixed. Simulation results show that the memory can operate down to 700mV and consumes on average 3.5μW at 0.8V. However, this is our first attempt with this technology, and thus we are uncertain about the accuracy of the simulation results.

This chip recently returned from fabrication. Initial testing results show that the bootup sequence consumes ~18μW when VDDH is set to 1.8V and VDDL is set to 1.0V. The ULP-BUS was connected to the SoC to verify the startup sequence. A simple GPIO toggle program was developed for this procedure. The SoC was capable of starting up from FeAR once the available energy exceeded the bootup threshold. Initial measurement results also showed that the ULP-BUS can operate over a wide range of voltages (0.35V-1.0V).

6.4 Individual Contribution

This work was performed in collaboration with Christopher Lukas. My contributions to the chip include:

1. designing and laying out the Fe-RAM cell, sense-amplifier, and array architecture.
2. simulating and verifying the read/write functionality of the array.
3. laying out the arrays with their peripherals.
4. defining the interfaces to the SoC and the ULP-BUS (in collaboration with Christopher).
5. helping with the top level integration of the chip (in collaboration with Christopher).

6.5 Conclusion

This chapter presented a ferroelectric auto-recovery (FeAR) sub-system as a back-up non-volatile memory for battery-less SoCs. A combination of power saving features were introduced into this sub-system to keep it within the harvesting power budget. A differential bit-cell with a reference-less sense amplifier improves the reliability of the memories and enables low voltage operation. An ultra-low power bus (ULP-BUS) was also implemented to enable on-package integration of FeAR with the SoC and at the same time improve the programming throughput and reduce the programming time. These features result in a non-volatile platform that requires $\sim$18μW for re-programming upon a power-on reset.

Chapter 7

Ultra-Low Power Always-on Voice Activity Detector

As illustrated in Section 1.1, the analog and digital components consume more than 60% of the total power within the system. Thus, reducing the power and energy consumption of these components will provide a significant reduction in the overall system power/energy. An example application that requires an analog front-end and extensive digital processing back-end is voice activity detection and processing. In this part of the research, architectural and circuit techniques will be investigated to minimize the power and energy consumption of a voice activity detector (VAD) while providing a good level of accuracy. We start by presenting a brief background of speech and how it is created before presenting the implementation.

7.1 Background

The different mechanisms used to create speech result in three classes of sounds: voiced, unvoiced, and plosive [41]. Voiced sounds (e.g. the letter a) are generated when airflow from the lungs causes the vocal tracts to oscillate and there is no constriction within the voice track. Since the vocal tracts are vibrating, the resulting speech signal appears to be quasi-periodic. On the other hand, unvoiced sounds (e.g. the letter f) are generated by

airflow passing through a constriction within the voice track. In this case, the vocal tracts are not vibrating and thus the generated speech signal appears to be almost random. Finally, plosive sounds (e.g. the letter p) are generated by building up pressure behind a complete closure in the voice track. When the closure opens, the pressure is released, and a brief sound is generated. While voiced and unvoiced sounds are easily distinguishable through a study of their time-domain signals or their frequency spectra, differentiating between unvoiced sounds and plosives is not a simple task, and thus the two are considered one class.

The speech signal has a wide frequency range (300 to $> 10KHz$). However, the spectrum drops off significantly at high frequencies. For voiced sounds, the peak of the spectrum is usually below 1KHz and frequencies above 4KHz can be ignored. However, for unvoiced sounds, the spectrum is more spread out and frequencies as high as 8KHz still contain some data. Thus, to accurately represent the speech signal digitally, sampling rates as high as 40KHz might be needed. However, a lower sampling rate can be used depending on the application. For voice activity detection, a sampling rate of 8KHz is chosen. A low pass filter is generally recommended before sampling for two reasons: to get rid of high frequency components that exceed the Nyquist frequency and to filter out any noise that might alias the signal [41].

When examined in time segments of 5-100ms, the speech signal appears almost invariant [41]. Thus, many algorithms rely on analyzing the speech signal and extracting different features over a short-time period usually referred to as a frame. Features include amplitude of the speech signal, short-time energy (the sum of the squares of the speech samples of a frame), pitch (the fundamental frequency of the signal), zero-crossing rate, formant frequencies (the vocal tract vibration frequencies that pass the most energy), and others. Different algorithms use different features or combinations of features to detect words or emotions. However, almost all algorithms follow the sequence in Figure 7.1. First, the speech signal is acquired through a microphone then passed through the analog front end where it is filtered, amplified, and digitized. Next, a transformation or a set of transformations is applied to the signal over

a windowed interval (frame) to extract the desired features. Finally, the extracted features are compared to different thresholds, and a decision is reached. To improve the accuracy of the decision when the signal-to-noise ratio is low, many algorithms adapt the thresholds and/or estimate and suppress the noise from silent frames.

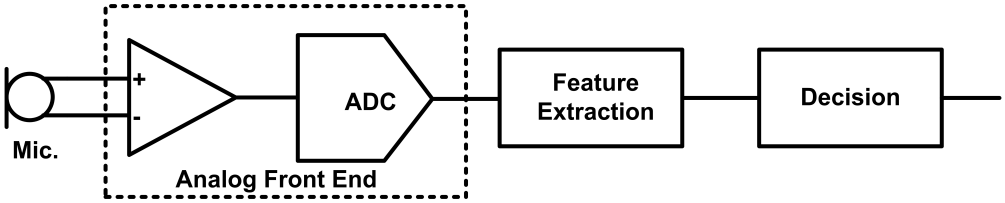


Figure 7.1: The speech pipeline.

7.2 Wake-up VAD Architecture

A ULP voice activity detector is developed to provide a sub-100nW wake-up signal to the battery-less SoCs. Since the zero-crossing algorithm is simple and computationally inexpensive, it was chosen for this implementation. Figure 7.2 shows the architecture of the wake-up VAD. A ULP continuous time comparator[1] takes an input signal from a microphone and determines whether the signal is above or below zero. The result is then fed into a digital block that keeps track of the number of zero-crossings that occurred within a time frame. Once that number exceeds a predefined/precomputed voice activity threshold, an interrupt signal is asserted. A SPI interface allows the SoC to program the different parameters of the implemented algorithm as well as the thresholds.

7.2.1 ULP Comparator

Figure 7.3 shows the low power comparator along with its support circuitry. The audio signal enters a self-biasing ULP continuous time comparator (0.25V - 3.3V operation) for zero detection. The bias generator consists of two long channel, reverse biased NMOS devices with

[1]The building blocks of this comparator were developed by Christopher Lukas.

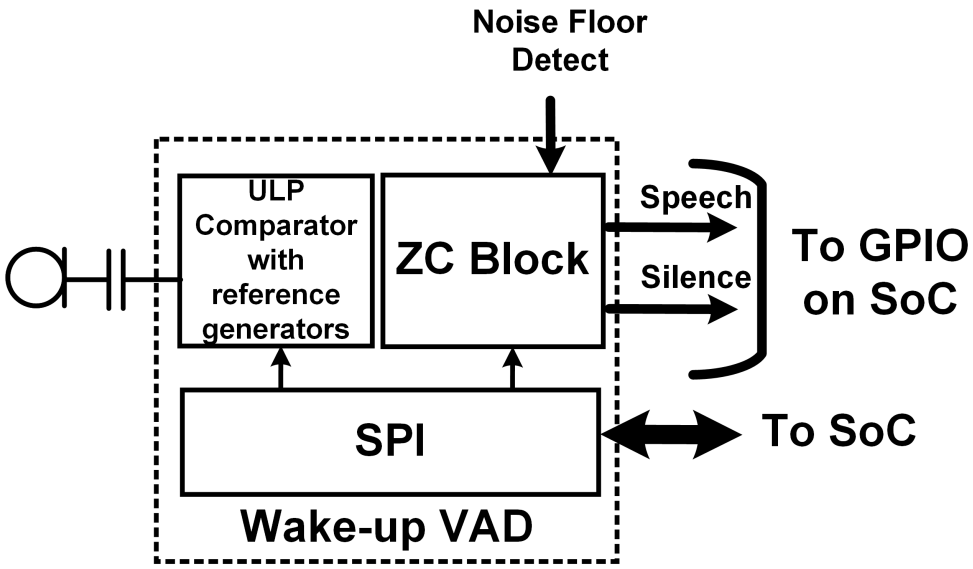


Figure 7.2: Block diagram showing the VAD architecture and its interface to a low power SoC.

two more digitally controllable transistors to improve the range of generated voltages. The comparator utilizes a current mirror, creating a single ended full-swing output appropriate for use in digital circuits. Due to the single-ended input potentially having no DC offset, one side of the differential input is tied to an off-chip series capacitor and another configurable bias generator to create such an offset if necessary. The other input is tied to a third configurable bias generator with the same configuration as the previous to create a voltage for zero-crossing comparison. Figure 7.4 shows a simulation of the comparator when provided with a 1KHz sine wave input centered around 800mV (typical of low power microphones). Simulation results show that the comparator consumes 1.2nW under these conditions.

7.2.2 ULP ZC Block

The output of the comparator is fed into a digital block implementing the zero-crossing (ZC) algorithm [41]. The algorithm was adapted to detect speech in real-time. This algorithm keeps track of the number of ZC within a sequence of frames. The frame size is usually between 5-100ms long. The frames are dispersed in time but usually overlap to avoid losing

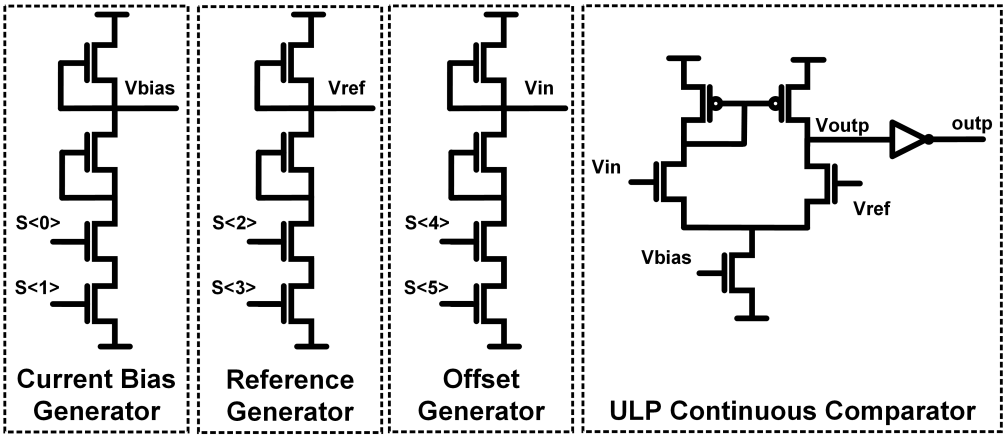


Figure 7.3: Circuit level description of the proposed ULP comparator.

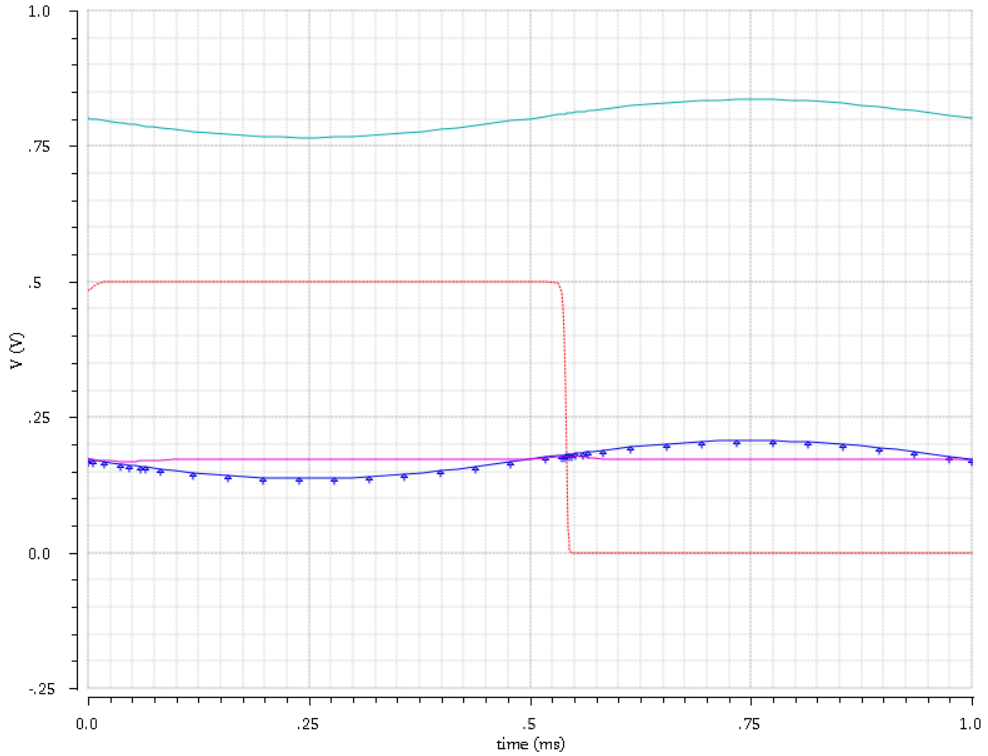


Figure 7.4: Simulation result showing the functionality of the ULP comparator. Green: microphone input, Blue: comparator input, Pink: reference input, Red: comparator output.

data. At the end of each frame, the number of ZC is compared to either a predefined or pre-calculated value, and a decision is made on whether speech was detected or not. To improve the accuracy of the algorithm, five consecutive frames must determine the presence of speech before the 'Speech' output is asserted and similarly for the 'Silence' output. The size of the frame can be programmed by the user, and the amount of overlap is always calculated

as one-fourth the frame size.

The main concept behind the ZC algorithm is that background noise will cause the number of ZC to increase whereas voice activity will reduce the number of ZC. To avoid using a predefined threshold and allow the algorithm to adapt to different environments, a noise floor detect is implemented to determine the background noise level and adapt the threshold. This procedure must be enabled externally when voice activity is absent. The algorithm calculates the number of ZC in a 300ms interval and determines the threshold.

To reduce the power consumed by this digital block, the number of hardware resources is reduced through serialization. The ZC algorithm relies on simple incrementer circuits to compute the number of ZC within each of the four frames. By noticing that a maximum of 8KHz sampling rate is required and the SoC can provide a 32KHz clock signal to the digital block, one incrementer can be used along with a multiplexer instead of four incrementers. Figure 7.5 shows a simulation result of the ZC algorithm processing a speech file with added white noise such that the signal-to-noise ratio is 15. The dashed blue line represents the 'Speech' output; whereas, the dashed green line represents the 'Silence' output. The figure shows that the implemented algorithm is capable of detecting different uttered words with high accuracy. The noise floor detect algorithm was used here to adapt the threshold of voice activity detection.

7.2.3 Short-Time Energy

Another VAD time-based algorithm was also implemented on this chip. The short-time energy (STE) measures the energy of speech signal within a frame. The STE algorithm provides higher accuracy than the ZC algorithm but requires a digitized version of the microphone input. In this implementation, a 12-bit digital input was fed into the digital block (no AFE or ADC were implemented). As with the ZC algorithm, a noise floor detect was implemented to determine the threshold of voice activity. Hardware reuse and serialization reduce the power consumed by this block. Figure 7.6 shows a simulation result of the STE algorithm

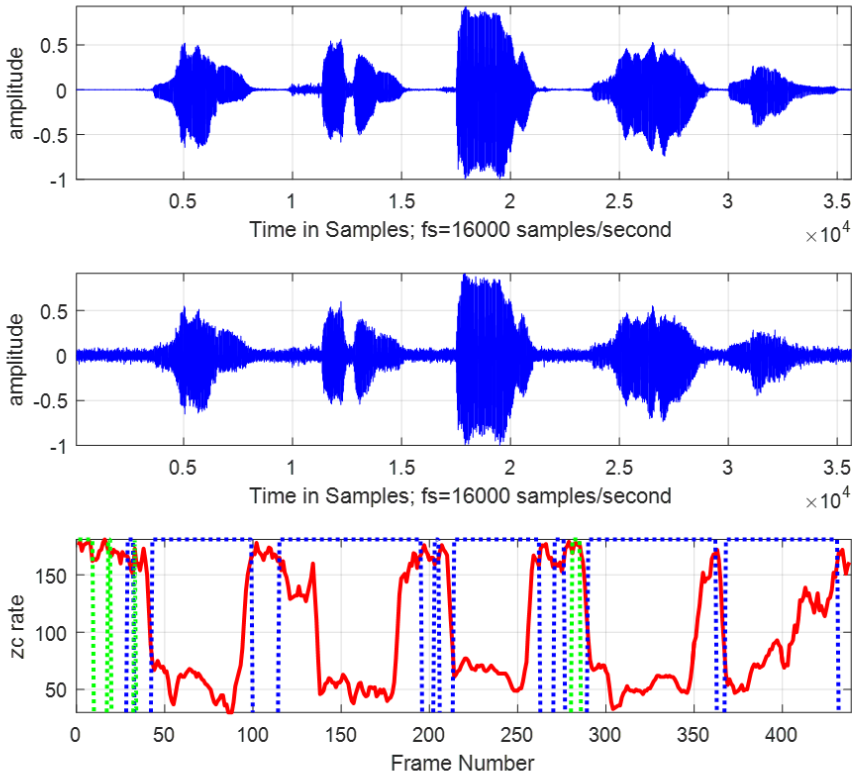


Figure 7.5: Simulation results showing the functionality of the ZC algorithm. Dashed blue line represents 'Speech', dashed green line represents 'Silence'.

processing a speech file with added white noise such that the signal-to-noise ratio is 15. The dashed blue line represents the 'Speech' output; whereas, the dashed green line represents the 'Silence' output. The figure shows that the implemented algorithm is capable of detecting different uttered words with high accuracy. The noise floor detect algorithm was used here to adapt the threshold of voice activity detection.

7.3 Chip Results

This proposed VAD was fabricated in a commercial 130nm process. Two revisions of this block were implemented. In the first revision, only the digital ZC and STE algorithms were taped-out. Figure 7.7 shows a die photo of this revision. The functionality of the ZC

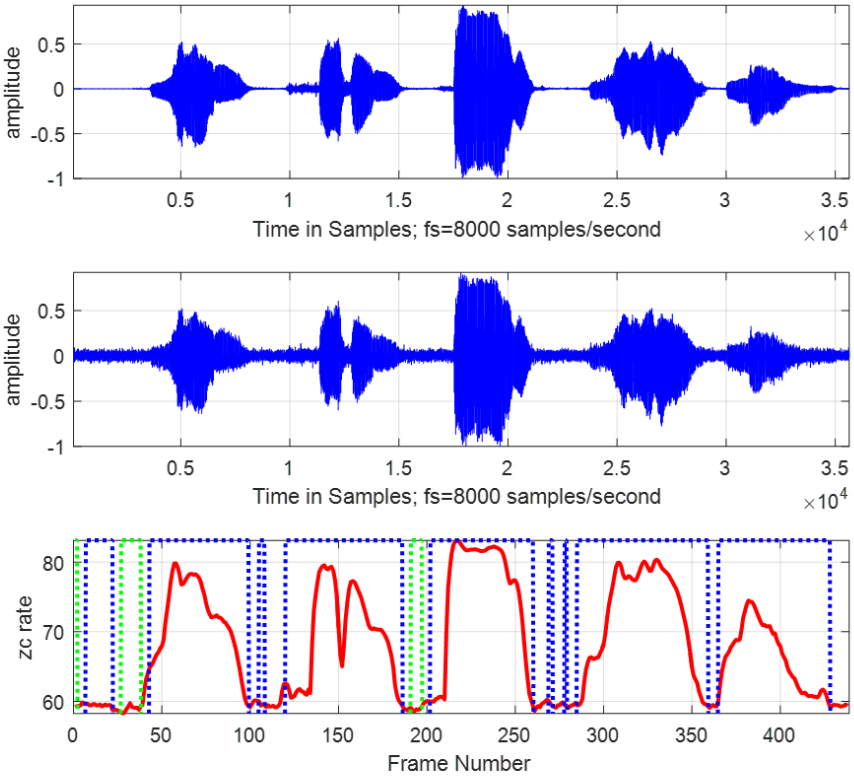


Figure 7.6: Simulation results showing the functionality of the STE algorithm. Dashed blue line represents 'Speech', dashed green line represents 'Silence'.

algorithm was verified at 0.5V with 32KHz clock. This algorithm consumes only 4.5nW, which is significantly lower than that of previously published high-accuracy implementation [42]. The second version includes the comparator and the SPI interface. However, it is still in fabrication.

7.4 Conclusion

In this chapter, an ultra-low power always-on voice activity detector for battery-less systems was developed. The implemented VAD relies on a zero-crossing algorithm implemented with a ULP continuous comparator. A noise floor detect algorithm allows the VAD to adapt to background noise and avoid false positives. The digital implementation relies on hardware

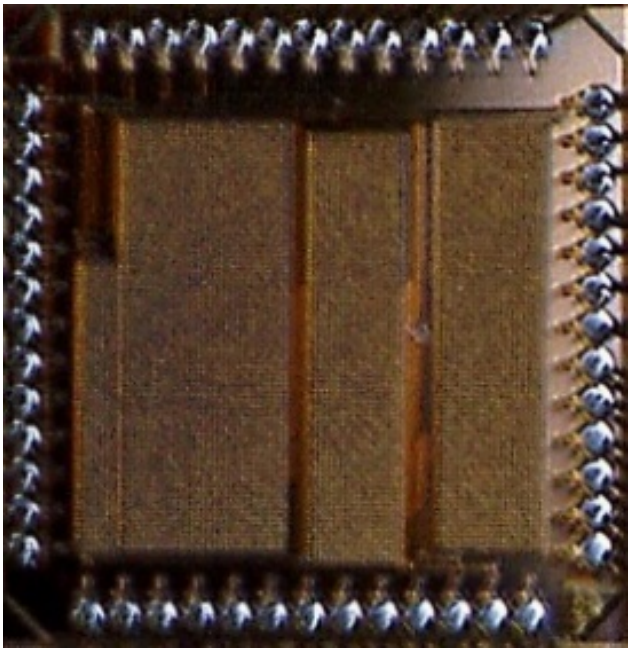


Figure 7.7: Die photo of the first version of the chip.

reuse and serialization to reduce the power consumption down to 4.5nW, making it ideal for battery-less system wake-up. A short-time energy algorithm is also implemented as a higher accuracy algorithm but consumes more power.

Chapter 8

Sub-μW Battery-less SoC

In the previous chapters, we described techniques to improve the reliability and reduce the power consumption of on-chip and off-chip memories and a specialized sensing interface. In this chapter, we integrate these blocks with an energy harvesting platform power manager and other sensing interfaces to achieve a completely self-powered system-on-chip (SoC) for IoT applications. The SoC is designed as part of a system in package (SiP) including a radio (FSK) transmitter and non-volatile memory (NVM) (Chapter 6). The different features in each of the previously developed blocks are leveraged in the SoC to significantly reduce power consumption, improve reliability, and enable recovery.

8.1 System Architecture

Figure 8.1 shows a block diagram of the SoC and its interfaces to off-chip on-package communication and auto-recovery sub-systems. The SoC consists of an energy harvesting platform power manager (EH-PPM), three sensing interfaces, a custom low power controller (LPC), a suite of hardware accelerators, an on-chip crystal oscillator, and a power monitor (PM) with a cold-boot management system (CBMS) that enables recovery from FeAR after complete power loss. The SoC also interfaces with an FSK transmitter to communicate gathered data to a remote base station for further processing.

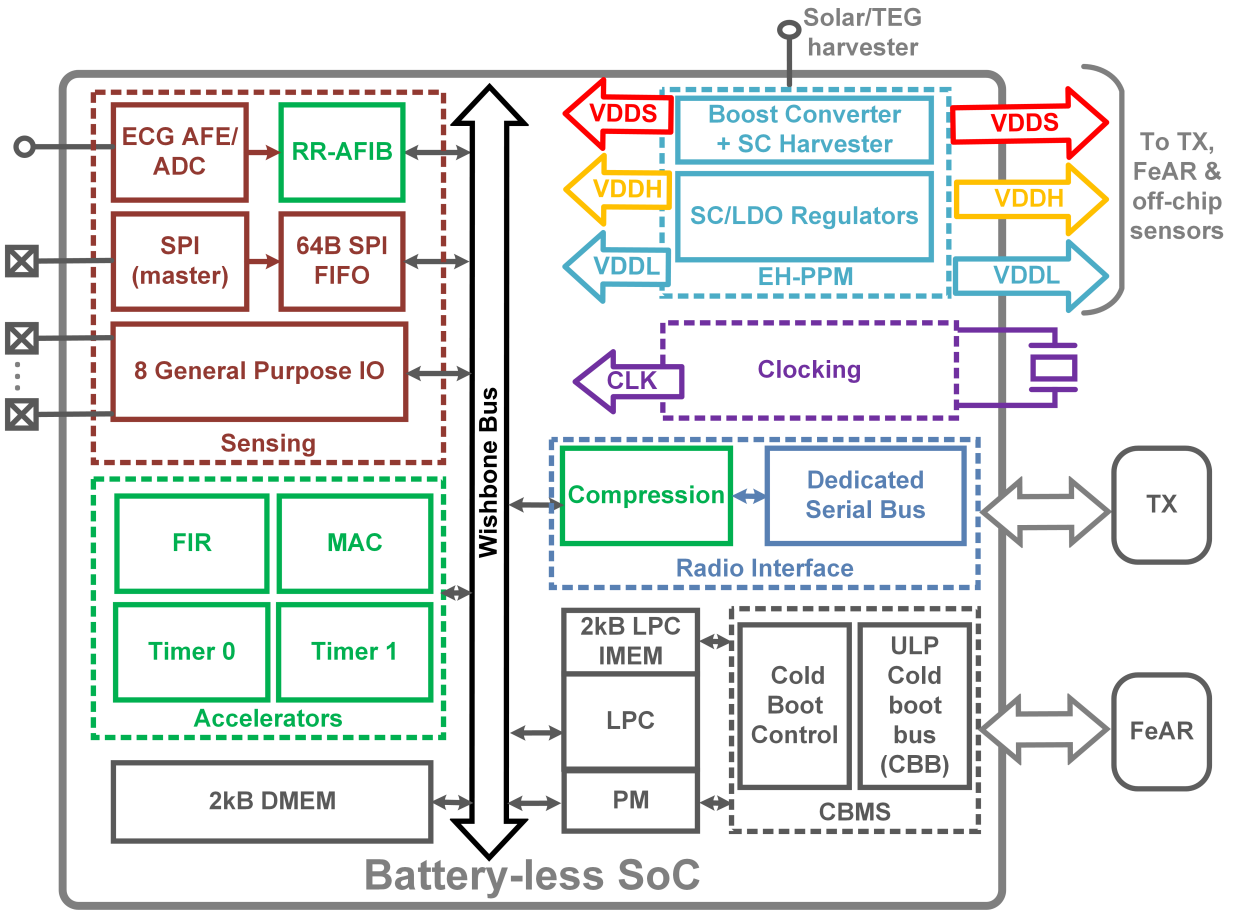


Figure 8.1: Block diagram of the battery-less SoC and its interface to the radio chip and FeAR.

8.1.1 Energy Harvesting Platform Power Manager (EH-PPM)

The EH-PPM[1] (Figure 8.2) is responsible for powering the various components in the SiP. It harvests from either Photovoltaic (PV) or Thermoelectric Generators (TEG) using either a single-inductor boost converter with maximum power point tracking (MPPT) control [43] or a fully integrated (no external passives) voltage doubling switched-cap harvester. The harvested energy is stored on a 10mF supercapacitor, and an on-chip clamp circuit limits the maximum voltage of the supercapacitor, V_{CAP}, to 1.5V to prevent device damage. Under favorable indoor harvesting conditions, a PV cell with an open circuit voltage, $V_{OC} > 1.2$V, can directly charge the supercapacitor, thus bypassing the boost converter completely. The EH-PPM

[1]This unit was developed by Abhishek Roy.

also consists of three fully integrated regulators that deliver three voltages: a sub-threshold voltage rail (0.5V) for the main control, a nominal voltage rail (1V) for the SoC pads and the on-package sub-systems, and a high voltage rail (1.8V) to power the sensing interfaces and any commercial-of-the-shelf (COTS) sensors. The regulators are specifically designed to handle sub-μW loads and use load-dependent pulse frequency modulated control and nW-power error amplifiers, comparators, and reference generators to reduce their quiescent current (~400nA).

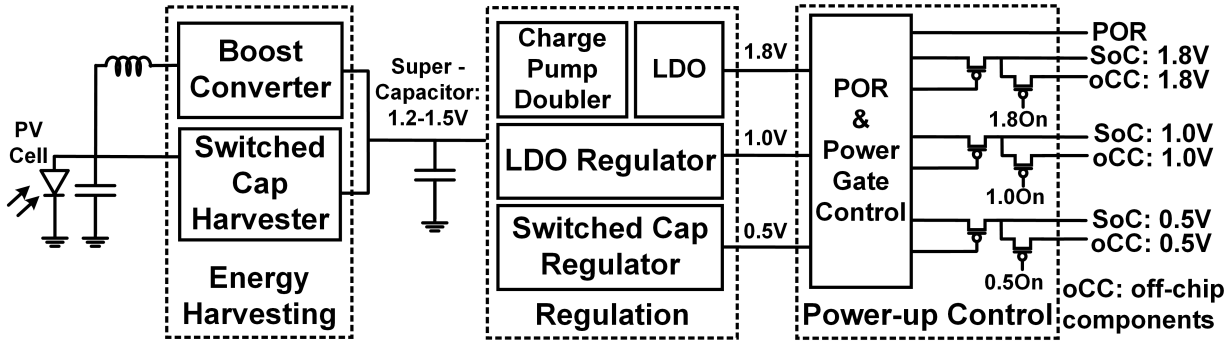


Figure 8.2: Main building blocks of the EH-PPM and its power-up circuitry.

A power-up circuit ensures a sequenced turn-on of the different regulators before enabling the system. Figure 8.3 shows the timing of the rail power-up sequence. To enable a smooth power-up for the rails, all major blocks in the SoC are either power-gated or held in standby mode. Since the SoC pads use two or more of the power rails provided by the EH-PPM, a power-on-reset (POR) feature is added to the pads. This feature ensures that the pads do not draw extra current when some of the rails are still turning on.

8.1.2 SoC Startup Sequence

After the EH-PPM startup sequence concludes, the POR is asserted and the SoC is loaded onto the three rails. Once power is supplied, the clocking block is the first block that starts up. It includes a 31.25 kHz off-chip crystal oscillator with a low-power on-chip oscillator [44]. The crystal startup sequence then follows to ensure the crystal oscillator is stable before the digital and sensing interfaces turn on. The crystal startup circuit consists of a simple four-second

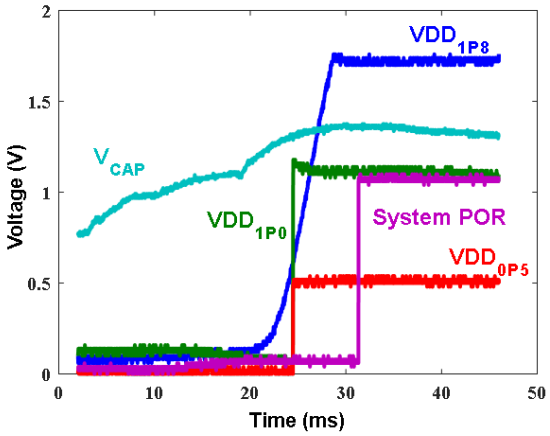


Figure 8.3: The power-up sequence of the EH-PPM showing V_{CAP} charging up above 1.2V and the rails ramping up.

counter that generates a crystal-on-reset (XOR). Once XOR is asserted, the PM handles the startup of the digital and sensing interfaces by keeping track of the available energy through monitoring V_{CAP}. As soon as V_{CAP} exceeds the bootup threshold, the PM signals the CBMS to start the bootup sequence. The CBMS sends a BOOTUP command to FeAR and programs the SRAM with the data it receives back. Once the SRAM is programmed, the PM asserts the system-on-reset (SOR) which starts the digital and sensing interfaces.

8.1.3 The Power Monitor (PM)

As mentioned before, the PM is responsible for monitoring the energy available to the system by measuring V_{CAP}. To do that, the PM relies on a voltage controlled ring oscillator (VCO) that translates V_{CAP} into a variable frequency clock signal. A counter is then used to count the number of edges within a programmable time period and infer V_{CAP} based on the sum. To reduce the power consumption of the system, the VCO is disabled for a programmable number of cycles. Based on the available energy, the PM will decide to turn on or shut off parts of the system. The PM is also responsible for managing the startup and shutdown sequences of the SoC. It has six main operating modes: IDLE, RED, YELLOW, GREEN, BOOTUP_WAITING, and BOOTUP_STARTED (Figure 8.4). On startup, the PM is enabled

and in IDLE mode. If the chip is coming from an XOR, the PM goes into BOOTUP_WAITING mode until V_{CAP} exceeds the boot-up threshold voltage (BOOTUP_THRESHOLD). Then, it transfers to BOOTUP_STARTED mode where it instructs the CBMS block to retrieve program and critical data from FeAR.

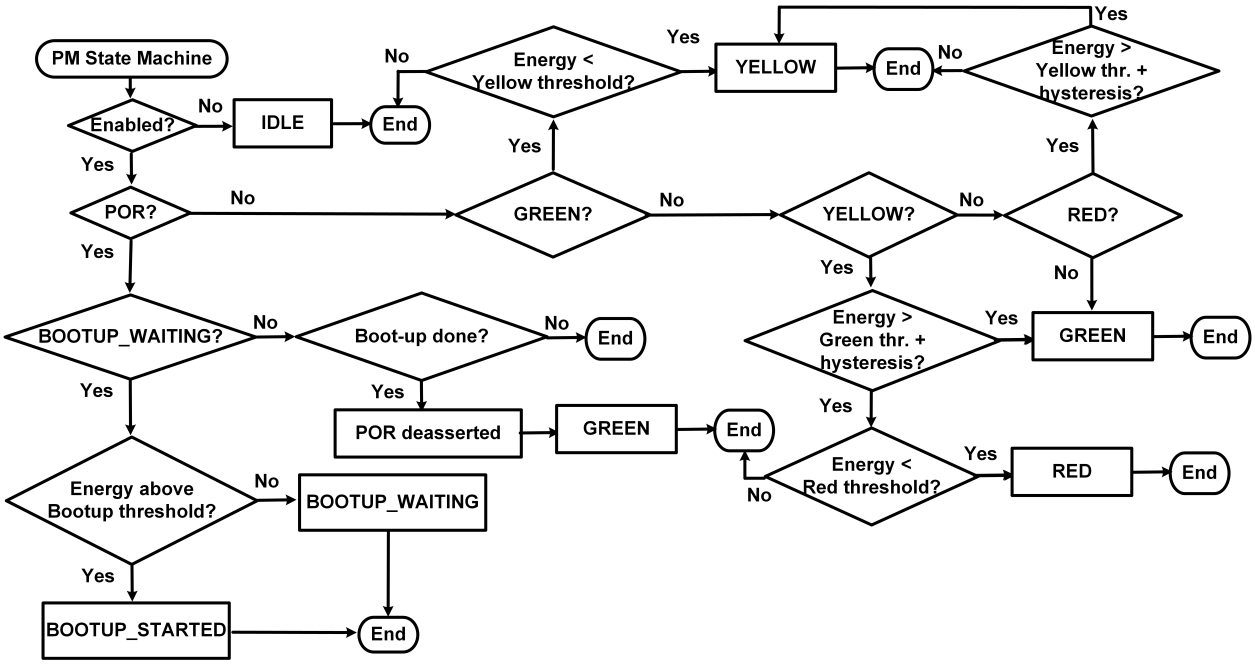


Figure 8.4: Flowchart showing the conditions under which the PM changes its state.

Once the SoC is programmed, the PM turns on the LPC and enters GREEN mode. Once in GREEN mode, the PM goes back to monitoring V_{CAP} voltage. If V_{CAP} drops below the YELLOW_THRESHOLD, the PM transfers to YELLOW mode and power/clock gates some of the blocks as defined by the user in YELLOW_POWER_SETTINGS. If V_{CAP} drops below the RED_THRESHOLD, the PM goes into RED mode and power/clock gates more blocks as defined by the user in RED_POWER_SETTINGS. Going from RED to YELLOW or YELLOW to GREEN, the V_{CAP} must exceed the appropriate threshold by the amount specified in the user-defined HYSTERSYS register. In each of the RED, YELLOW, and GREEN modes, the user can decide which blocks to power/clock gate to ensure that the system remains within the power budget. The user can also specify in which mode (RED, YELLOW, or GREEN) to run the system backup into FeAR. Once the PM enters that mode,

a backup signal is sent as an interrupt to the LPC. The LPC programmer is then responsible for transferring the critical data into the CBMS.

8.1.4 Cold-Boot Management System (CBMS)

The CBMS is responsible for programming the SoC from FeAR and backing up to FeAR any critical data the SoC produces. Figure 6.5 shows the bootup and backup sequences of the CBMS. The backup sequence (Figure 6.5a) is started when the PM detects a droop on V_{CAP} that might cause complete power loss. The PM then signals the LPC to collect any critical data and send it to the CBMS. Once the LPC completes this data transfer, the CBMS powers on FeAR and sends a BACKUP command followed by the data to be saved. After a power loss event, the PM starts the bootup sequence (Figure 6.5a). The CBMS powers on FeAR, retrieves data from the Fe-PROM first, and uses it to program the instruction memory on the SoC. Once the program EOF is reached, critical data is retrieved from the Fe-FIFO and saved within the CBMS for the LPC to retrieve when the system starts.

Even though FeAR is not always powered, integrating and managing it within a self-powered system budget is challenging because of the inherent high powered nature of existing non-volatile technology. Thus, a ULP on-package cold-boot bus (CBB)[2] interfaces between the SoC and FeAR. The CBB reduces FeAR on time through bus parallelization and power gate control. Figure 8.5 shows the reduced-swing driver on the SoC and the corresponding low-swing receiver on FeAR. Since the data transferred on the CBB does not need to be full-swing, the power consumed during transmission of data is reduced from CV_{DD}^2 to $CV_{DD} \times \Delta V$ where ΔV is the swing on the bus. The low-swing receivers are also designed to detect as low as 100mV swing on the bus (based on simulation results). The combination of power-saving techniques introduced on FeAR and the CBB will allow the integration of non-volatile capabilities into battery-less systems.

[2]The ULP CBB was developed by Christopher Lukas.

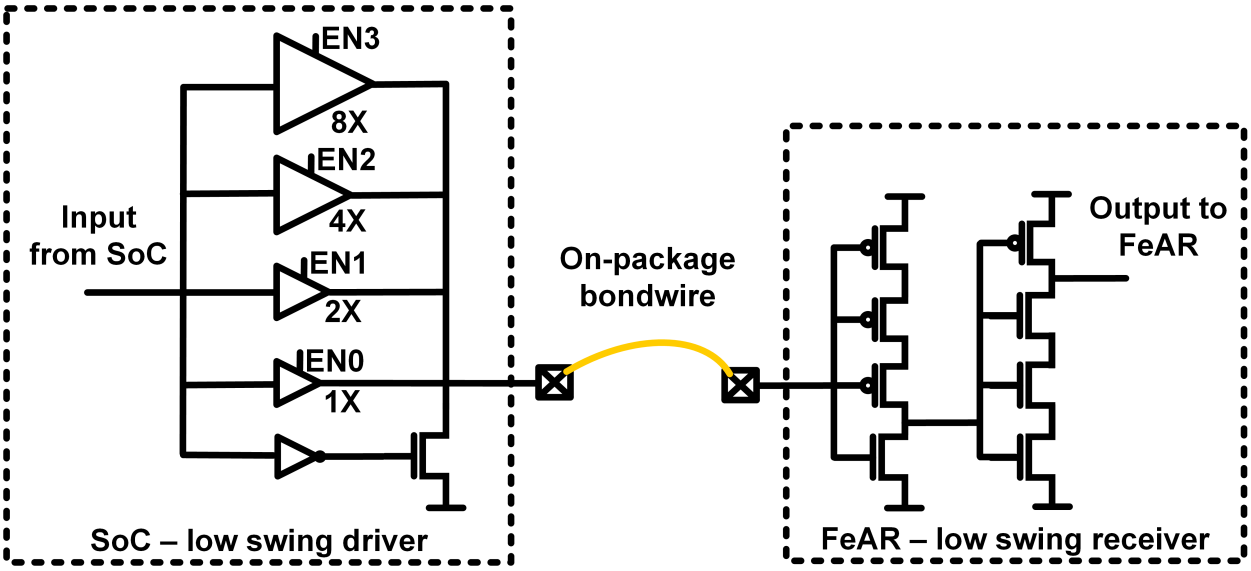


Figure 8.5: The low swing driver and receiver within the CBB.

8.1.5 The Low Power Controller (LPC) and its Instruction Memory (IMEM)

The LPC is adapted from [2] to include an arithmetic/logic unit (ALU) and take advantage of the different features within the instruction SRAM memory. The SRAM is specifically designed for self-powered systems with its high-threshold voltage 8T bit-cell and myriad power saving techniques (Chapter 4) and is tightly coupled with the LPC. The read burst mode in the SRAM reduces the power of the control block (LPC + SRAM + Radio Interface + PM + CBMS) by up to 17%. The standby mode of the SRAM is linked to the stall state of the LPC. In a stalled state, the power consumption of the LPC and instruction memory is reduced by up to 61% compared to active state. For program sizes below 1KB, users can further reduce the power consumption ($\sim$23%) of the control block by disabling the unused bank of the memory. Combining the three features together allows up to 65% reduction in the control block power. Figure 8.6 summarizes the measured power reduction due to the tight coupling between the LPC and SRAM.

The LPC instruction set and the available accelerators target IoT applications and thus allow users to develop compact programs without sacrificing functionality. Table 8.1 shows

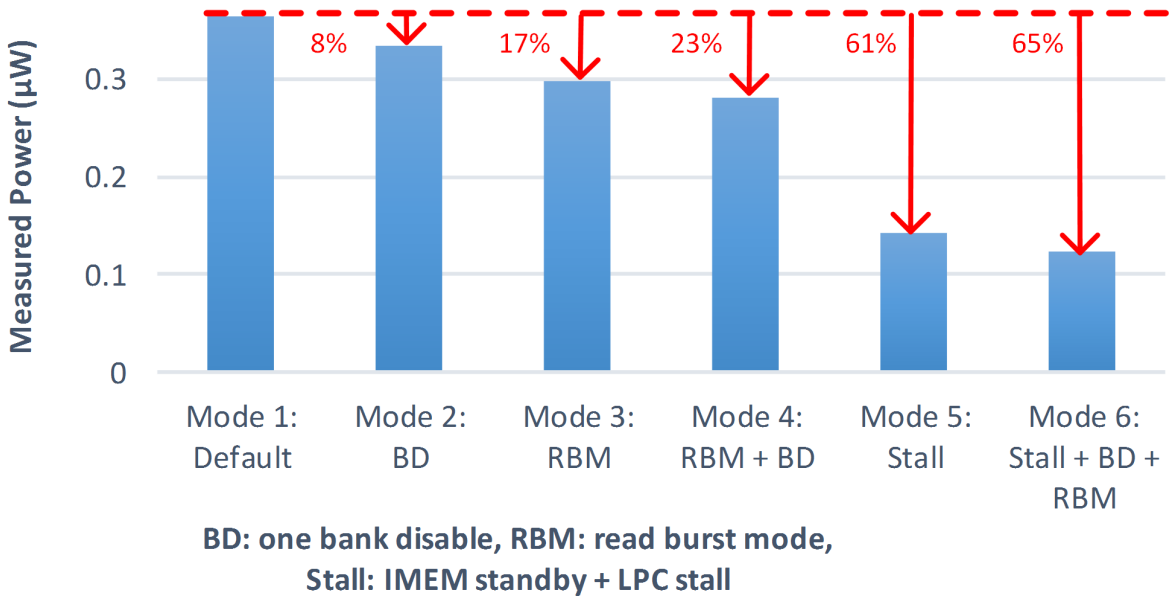


Figure 8.6: The SRAM power modes reduce the measured power consumption of the control block.

the instruction set of the LPC. A python-based assembler was developed to translate assembly style instructions into Verilog test-benches to aid in testing and verification of the SoC.

8.1.6 The Accelerators

The SoC includes a number of accelerators that target ULP, low throughput applications. These include a multiply-accumulate (MAC) unit, finite impulse response (FIR) filter, two timer units, a compression block, and a heart rate (RR) and atrial fibrillation (AFIB) block. To reduce the system complexity without compromising the level of integration, the accelerators needed by any sensing or communication interface are directly integrated with their corresponding interface. This architectural choice allows for a one-bus system without a direct memory access (DMA) interface and without loading the LPC with data transfers between blocks. Thus, the RR-AFIB block is integrated with the ECG analog front end (AFE) and the compression block is integrated with the radio interface. Following is a brief description of each accelerator.

Table 8.1: The LPC instruction set along with a description of each command.

Instruction	**#cycles**	**Description**
NOP	1	No operation
BUSW	1	Write into a memory mapped register on the bus
BUSR	2	Read from a memory mapped register on the bus
MOVL	2	Move a literal into a memory mapped register or the LPC registers
JMP	1	Unconditional Jump
CJMP	1	JMP EQ - Jump if LPC register 1 = register 2 JMP GE - Jump if LPC register 1 >= register 2 JMP RD_FULL - Jump if Radio FIFO is full JMP TMR0_EXP - Jump if Timer 0 expired JMP TMR1_EXT - Jump if Timer 1 expired JMP GPIO0_SET - Jump if GPIO 0 is set JMP GPIO1_SET - Jump if GPIO 1 is set JMP RD_TX_DN - Jump if Radio transmission is done JMP RD_SPI_DN - Jump if SPI transmission to Radio is done JMP RR_HR_DN - Jump if the heart rate measurement is done JMP RR_AFIB_DET - Jump if afib is detected JMP CBMS_BCKUP - Jump if the backup flag is asserted
STALL	1	STALLP - stall until LPC IMEM is programmed STALLJ - stall until JTAG is disabled STALL TMR0_EXP - stall until Timer 0 expired STALL TMR1_EXT - stall until Timer 1 expired STALL GPIO0_SET - stall until GPIO 0 is set STALL GPIO1_SET - stall until GPIO 1 is set STALL RD_TX_DN - stall until Radio transmission is done STALL RD_SPI_DN - stall until SPI transmission to Radio is done STALL RR_HR_DN - stall until the heart rate measurement is done STALL RR_AFIB_DET - stall until afib is detected
SAVEPC	1	Save the PC of the next instruction into LPC _PC_SAVED register
RESTPC	1	Restore PC - LPC_PC= LPC_PC_SAVED
ADD	1	Add two LPC registers
SUB	1	Subtract two LPC registers
AND	1	Bitwise AND of two LPC registers
OR	1	Bitwise OR of two LPC registers
XOR	1	Bitwise XOR of two LPC registers
NOT	1	Bitwise NOT of an LPC register
SHIFTL	1	Shift the contents of an LPC register left by one position
SHIFTR	1	Shift the contents of an LPC register right by one position
ROTL	1	Rotate the contents of an LPC register left by one position
ROTR	1	Rotate the contents of an LPC register right by one position
STAT	1	Gives access to internal registers: Interrupt Enable register, Interrupt PC register, Interrupt Flag register, IMEM Configuration register

The **compression**[3] accelerator implements the differential entropy compression algorithm, a loss-less algorithm designed to compress raw sensor data by exploiting the temporal correlation between consecutive samples. Since the main function of this block is to reduce the communication power by reducing the number of transmissions required, it is integrated directly with the radio interface. Thus, it takes the sensor data required for transmission from the bus, and sends the compressed data directly to the FIFO within the radio interface.

The **RR** block [2] implements a simplified version of the Pan-Tomkins algorithm [45] to calculate the R-R interval. The RR block takes its input from the 12-bit analog-to-digital converter (ADC) and calculates the heart rate. Once the heart rate is detected, it sends a signal to the LPC which can be used to interrupt the LPC or to recover it from a stall condition.

The output of the RR block is also used by the **AFIB** block [2] which implements the algorithm defined in [46]. The AFIB block keeps track of the number of AFIB events detected and the last 12 R-R intervals detected.

The **FIR** block [3] is a four-channel, 16-bit filter, with each channel having up to 16-taps. The number of coefficients, number of active channels, and number of parallel filters are programmable. Each channel can be independently clock gated when not in use.

The two **timer** blocks [3] include both counter and capture/compare features. In counter mode, they can be programmed to increment, decrement, or rollover, and they include a clock divider for increased range. Each timer generates an interrupt to the LPC that can also be used to retrieve the LPC from a stall state.

The **MAC** unit can be configured for multiply or multiply-accumulate feature. It takes two 16-bit signed inputs and provides a 32-bit output with an overflow flag.

[3]This unit was developed by Jacob Breiholz.

8.1.7 The Sensing Interfaces

The SoC includes three sensing interfaces to collect data: an ECG analog front end (AFE) with a 12-bit ADC, a serial peripheral interface (SPI) master with a 64-byte FIFO, and eight general purpose input/output pads (GPIO). The AFE[4] is implemented with an AC coupled non-chopper instrumentation amplifier operating in the sub-threshold. It takes a differential ECG signal as an input and produces a single-ended output that is sampled by the ADC. The AFE consumes 68.5 nW of power and has a programmable mid-band gain of 31-52 dB with tunable low pass and high pass corner frequencies. The low pass corner frequency is tunable from 40-155Hz. The ADC is a successive approximation (SAR) ADC that contains a split capacitor bank and a ground referenced comparator. Both the AFE and ADC operate at sub-threshold voltages.

The SPI interface with its FIFO allows the SoC to configure and communicate with COTS sensors. The SPI pads[5] include a custom specially designed level shifter with diode connected transistors to enable flexible operation between 0.4V and 3.3V. This feature will enable the SoC to efficiently communicate with ultra-low power research-based sensors while retaining compatibility with COTS sensors.

Eight GPIO pads [6] are also included in the SoC. These pads can be configured as inputs or outputs, and they use the same level converters in the SPI pads to communicate to research-based and COTS sensors alike. They are also equipped with a weak pull-down path to avoid spurious currents if the pad is not driven in the input mode. Two of these pads are connected as interrupt sources to the LPC and can be used to recover the LPC from a stall state.

[4]This unit was developed by Avish Kosari.
[5]These pads were developed by Christopher Lukas.
[6]These pads were developed with the help of Divya Akella and Christopher Lukas.

8.2 Chip Results

The SoC was fabricated in a 130nm commercial technology. Figure 8.7 shows the die photo. The main functionality of each of the building blocks was verified, and an example application was developed to highlight the advantages of the proposed SoC.

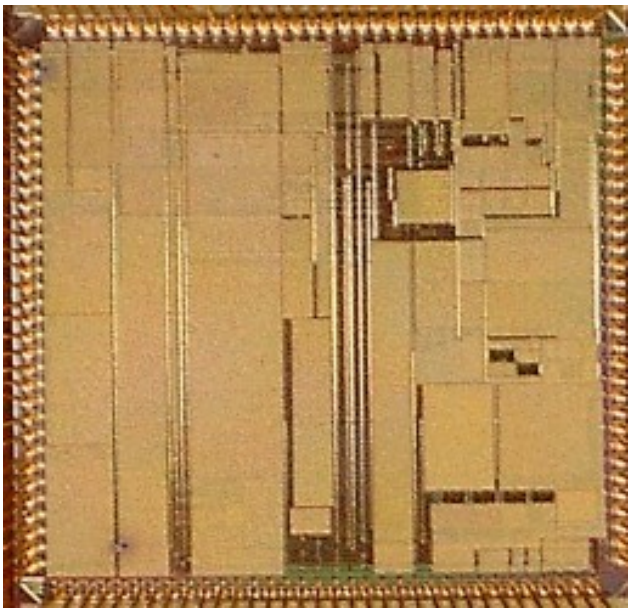


Figure 8.7: Die photo of the fabricated SoC.

In the example application (Figure 8.8), the SoC configures a COTS accelerometer [47] for free-fall detection and data logging. When a free-fall event occurs, the SoC wakes up, reads the data log, and compresses it before transmitting the data to the radio chip. In this application, the SPI and GPIO sensing interfaces are utilized, the LPC low power stall state reduces the power consumed by the system, and the compression block reduces the required number of transmissions.

Figure 8.9 shows the measured power distribution during different operating conditions. The GPIO and SPI sensing interfaces contribute most to the power consumption since they operate at 1.8V to communicate with the accelerometer. During the stall state, the SPI interface is completely power gated to reduce its power consumption.

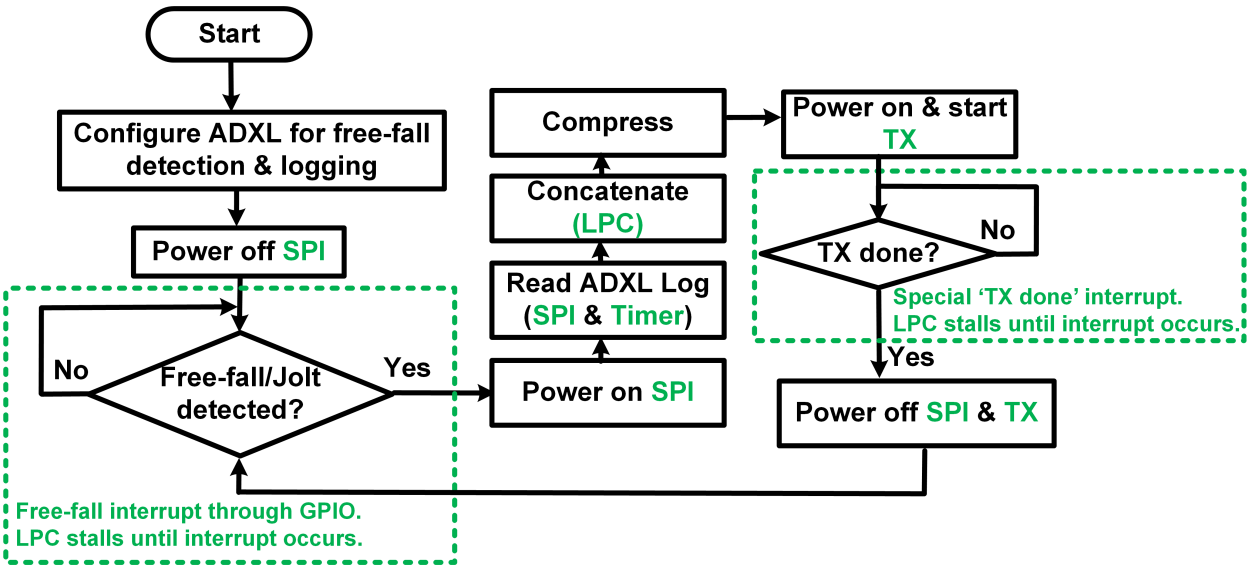


Figure 8.8: Die photo of the fabricated SoC.

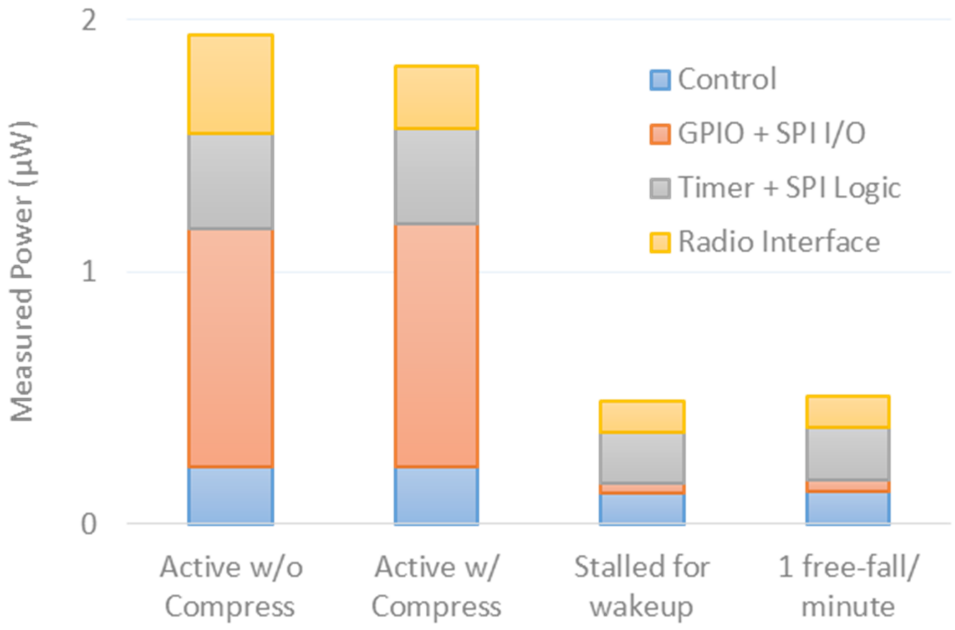


Figure 8.9: Measured power distribution of the different SoC building blocks during the different operating modes.

Finally, Table 8.2 highlights the main features of the SoC and compares it to similar state-of-the-art SoCs. Compared to similar systems, our SoC consumes 2x less active power and up to 10x less total power compared to similar SoCs [3].

Table 8.2: Comparison to state-of-the-art SoCs. MCU: main controller + instruction memory.

	This work	[48]	[49]	[8]	[3]
Battery-less	**Yes**	No	No	Yes	Yes
Harvests power	**Yes**	No	Yes	Yes	Yes
Fully integrated EH-PPM	**Yes**	No	No	No	No
Powers off-chip Sensors	**Yes**	No	No	No	No
Regulated voltages	**1.8V, 1.0V, 0.5V**	0.25V-1.2V	-	unregulated	1.2V, 0.5V, variable
Interface to NVM	**Yes**	No	No	No	No
On-chip SRAM	**4KB**	24KB	3.7KB	256B	12KB
Accelerators	**5**	2	3	-	7
Sensing interfaces	**3**	2	1	-	2
Total Power	**507nW**	850nW	45nW	295 pW	2.3W
Components Included in Total Power	**MCU + SPI + IO + Timer + GPIO + RI**	MCU	AFE + DSP	MCU	MCU + IO + SPI + FIR

8.3 Individual Contribution

This work was performed as part of a team including Christopher Lukas, Abhishek Roy, Jacob Breiholz, Harsh Patel, Ningxi Lui, Divya Akella, Shuo Li, Xing Chen, Avish Kosari, and Oluseyi Ayorinde. Christopher Lukas and I co-led this project, and my contributions to the chip include:

1. defining the different specifications of the SoC (with Christopher Lukas).
2. defining the system architecture of the SoC (with Christopher Lukas).
3. defining the interfaces between the digital blocks (with Christopher Lukas).
4. defining the power on sequence of the SoC (with Abhishek Roy and Christopher Lukas).
5. designing the on-chip memories (with Harsh Patel).
6. designing the Radio interface (with Christopher Lukas and Xing Chen).
7. designing the FeAR interface (with Christopher Lukas).

8. updating the LPC to include the arithmetic and logic unit and to take advantage of the new features of the SRAM.

9. updating the PM to include bootup and backup states.

10. designing the crystal startup sequence (with Divya Akella and Christopher Lukas).

11. implementing the ADC (with Aatmesh Shrivistava, Christopher Lukas and Ningxi Lui).

12. developing a python based assembler to help with the verification and testing of the SoC.

13. verifying the complete digital block (with Christopher Lukas and Jacob Breiholz).

14. integrating the different blocks in the SoC tapeout (with Christopher Lukas).

15. testing the SoC dice (with the rest of the team).

8.4 Conclusion

In this chapter, a 507nW self-powered SoC was demonstrated for ULP IoT applications. The SoC includes ULP SiP interfaces that enable its integration with a radio transmitter (TX) and a non-volatile memory (FeAR). The energy harvesting platform power manager (EH-PPM) powers the SoC as well as off-chip components and is optimized for low quiescent power. It supplies the SoC with 0.5V, 1.0V, and 1.8V and can also power ULP sensors and the SiP components while running an example free-fall detection algorithm. A power monitor (PM) cold-boots the SoC from NVM and adapts the system's power consumption. The tight integration between the SoC's blocks enables sub-μW operation.

Chapter 9

Conclusion

Battery-less systems promise to enable self-powered IoT devices for health, environmental, and structural monitoring as well as many other applications that track signals having low-to-medium sampling rates. However, these systems face significant challenges that can be divided into three categories: small power budget, varying power budget, and reduced reliability. In this dissertation, a number of techniques were introduced to reduce and adapt the power consumption of the system and improve the reliability of its building blocks. We mainly address these challenges as they relate to memory arrays, whether volatile or non-volatile.

9.1 Summary of Contributions

Low Power 6T SRAM

- A combination of read and write assist techniques was introduced to reduce the V_{MIN} of SRAM arrays down to near/sub-threshold voltages.

- The proposed approach addresses write failures as well as row and column half-select failures.

- The proposed approach allows the most reduction in the array V_{MIN} as compared to previously proposed techniques

- Different read and write assist techniques were thoroughly evaluated to study their impact on both write and half-select. This study was performed for a wide range of supply voltages while taking into account process variations.

- The study highlighted the advantages of detecting the process corner of the chip and adjusting the applied assist accordingly.

Low Power 8T SRAM

- A 1KB SRAM chip was fabricated in 130nm CMOS targeting ultra-low power battery-less systems.

- The proposed array operates over a wide range of voltages between 350mV and 700mV.

- The proposed array addresses read/write failures using WL boosting as an assist technique and half-select instability using a read-before-write approach.

- The read energy is reduced through a read burst mode.

- Standby and Shutdown modes are introduced through aggressive power gating to reduce the power consumption of the array.

- The low power features within the array are accessible in battery-less systems and significantly reduce their total digital power (up to 65%).

Low Energy STT-RAM

- A low-V_T bit-cell is introduced to provide the high current required to ensure correct write operation into an STT-RAM.

- A programmable all-digital write driver is introduced to reduce the write energy of an STT-RAM array.
- A methodology to design an STT-RAM array with the proposed bit-cell and driver is presented.
- The proposed techniques allow up to 37% reduction in the write energy of STT-RAM arrays.

Low Energy Fe-RAM

- A ferroelectric auto-recovery (FeAR) sub-system was developed as a back-up non-volatile memory for battery-less SoCs.
- A differential bit-cell with a reference-less sense amplifier improves the reliability of the memories and enables low voltage operation.
- Circuits and architectural techniques were also introduced to reduce the energy consumed during read from FeAR.
- An ultra-low power bus (ULP-BUS) was also implemented to enable on-package integration of FeAR with the SoC and at the same time improve the programming throughput and reduce the programming time.
- FeAR is expected to consume in the few microwatts of power for every bootup operation and, thus, will be within the budget of an battery-less system.

ULP VAD

- An ultra-low power always-on voice activity detector for battery-less systems was developed in 130nm technology. The detector consumes 4.5nW when running at 0.5V and 32KHz.

- Zero-crossing and short-time energy algorithms were implemented to detect voice activity without consuming high power.

- A noise floor detect algorithm allows the VAD to adapt to background noise and avoid false positives.

- The digital implementation relies on hardware reuse and serialization to reduce the power consumption.

Battery-less SoC

- A completely battery-less 507nW SoC was presented with low power interfaces to non-volatile memory and radio communication chips.

- An energy harvesting platform power manager capable of powering the rest of the SoC as well as off-chip sensors was integrated into the SoC.

- Analog and digital sensing interfaces were designed to gather information from research-based sensors as well as commercial-off-the-shelf sensors.

- The low power 8T SRAM arrays were integrated as instruction and data memories. The controller of the system takes full advantage of the different features of the memory to reduce the power consumed.

- A power monitor with a cold-boot management system keeps track of the energy available to the system and recovers program and critical data from FeAR upon power loss.

- A compression block with a serial radio interface allows for low power communication to a remote base station.

9.2 Team and Individual Contributions

Much of the work presented in this dissertation was completed as part of a team. The initial version of the work presented in Chapter 3 was completed as part of a summer internship at ARM Inc. In that version, the study was performed on sub-20nm FinFET technology with the support and constant advice from my manager Vikas Chandra. The same analysis was then repeated at the University of Virginia with the support of Harsh Patel for a commercial 130nm technology to further support the findings.

The 8T SRAM presented in Chapter 4 was also developed as part of a team including Harsh Patel, James Boley, and Arijit Banarjee. My contribution to the chip included the design of the high-V_T bit-cell and its supporting assist circuitry and the verification and testing of that design.

The low energy STT-RAM presented in Chapter 5 was developed as part of a 1-year internship at Intel Corp. The problem statement was proposed by my manager Muhammad Khellah. Feedback and support was provided by different team members within the group as well as members of supporting groups. The model of the STT-RAM was also provided by supporting groups.

The auto-recovery system presented in Chapter 6 was developed with the help of Christopher Lukas. My contribution to the chip was the design of the memories and the control unit. I also participated in the definition of FeAR's interface to the SoC. The Ferroelectric technology was provided and supported by Texas Instruments. Special thanks to Steven Bartling, Sudhanshu Khanna, Wendy Barr, Hemalata Sangodkar, Scott Summerfelt, and Zakir Shaik from TI for their timely and constant support during the setup of the process design kit (PDK).

The low power voice activity detector presented in Chapter 7 was completed with the help of Christopher Lukas. My contribution to the chip was the development of the digital algorithm, the integration of the analog blocks, and the SPI interface to the SoC.

The SoC presented in Chapter 8 was a massive effort involving a large team of students:

Christopher Lukas, Abhishek Roy, Jacob Breiholz, Harsh Patel, Ningxi Lui, Divya Akella, Shuo Li, Xing Chen, Avish Kosari, and Oluseyi Ayorinde. I co-led the design of this chip and thus was closely involved in defining the specifications and architecture of this chip. My contributions to the chip also include designing the SRAMs, updating the LPC and the PM, and designing the interface to the radio chip and FeAR.

9.3 Open Problems

Low Power 6T SRAM Even though the proposed combination of assist presented in Chapter 3 reduces the SRAM array V_{MIN}, a more thorough study is required to determine the impact of these techniques on the power and energy consumption of the array. An initial study showed that the optimal write assist technique for reducing the SRAM power/energy is different from the one that reduces the SRAM V_{MIN}. The presented analysis also ignored the impact of generating the assist voltages on the total energy and power of the complete system.

The corner analysis study presented in Section 3.4 highlighted the advantages of having a process monitor. Thus, a more complex SRAM system can be developed that integrates an accurate process monitor with an assist controller to reduce the power/energy of the system as well as improve its robustness to variations. This feature is particularly attractive for sub-threshold designs where the impact of variation is high which causes low yield.

Low Power 8T SRAM The 8T SRAM array presented in Chapter 4 contains a myriad of techniques to improve robustness and reduce power. However, the approach to address half-select failures – read-before-write – increases the energy consumed during a write operation. Other techniques to address this issue – such as the techniques proposed in Chapter 3 – might result in lower power. During a read operation, the read bit-line is completely discharged and a simple inverter-based sense amplifier is used. While this approach eliminates the need for a

reference, it increases the energy consumed during a read operation. Techniques to further reduce the read energy must also be studied.

Low Energy STT-RAM While the proposed techniques reduce the write energy, they negatively affect the read reliability. In the read analysis, a standard sense amplifier was used and thus showed some degradation in the read margins. A smart read sense amplifier might help mitigate these effects and reduce the energy consumed during a read operation.

More accurate models of STT-RAM and more advanced STT-RAM configurations were developed since this work was presented. Re-evaluating the proposed techniques with more accurate models and other STT-RAM configurations will help further support the findings in this dissertation.

Low Energy Fe-RAM Since the Fe-RAM array developed was the first attempt for our group to use ferroelectric technology, a thorough analysis of the capabilities and limitations of this technology have not yet been explored. The available models for the ferroelectric capacitor did not include local variation information, and simulation results might not map well to silicon data. Thus, when the working chips are retrieved, it is important to check how well the models match the silicon.

Since FeAR was developed as a complementary sub-system, we tried to minimize the communication and SRAM programming overhead through the ULP-BUS. Another approach that could be explored is integrating the non-volatility into the SRAM cells within the SoC. This requires porting the design of the SoC into a technology that includes non-volatile elements. However, the advantages of such an approach could be analyzed and compared to the results obtained from FeAR to determine which of the two approaches will result in the lowest power without compromising the reliability of the saved data.

ULP VAD The voice activity detector can be further improved by introducing an analog front end (AFE) with an analog-to-digital converter (ADC). However, both these components

must be optimized for low power applications. Traditionally, an oversampling ADC is used to capture the microphone input with high signal-to-noise ratio. Alternative architectures more suitable for battery-less systems should be investigated to determine the tradeoff between accuracy and power consumption.

The application space of the proposed system can be easily expanded to include emotion detection, keyword detection, and wheezing (health condition). For emotion detection, the outputs of the zero-crossing and short-time energy algorithms are monitored and compared to different emotion thresholds. Training might be required to ensure higher accuracy for each individual. For keyword detection, the short-time energy can be used but might not provide high accuracy. A wheezing algorithm [50] based on short-time energy was recently developed and could be adopted to improve the application space of the proposed SoC.

Battery-less SoCs The power/energy consumption of the battery-less SoC can be further reduced by enabling the different blocks to operate at their minimum energy points. However, this requires a thorough analysis to determine the tradeoffs between the reduced digital/analog power consumption and the increased power consumption required to generate the different voltages and frequencies.

The power manager within the SoC can be upgraded to track the energy required to complete a particular task and the energy available to the system. This will allow the SoC to better adapt to variations in the power budget. The custom low power controller reduces the power consumption of the system but needs a compiler to enable users to more easily develop applications for it. To protect the privacy of the SoC users, a security accelerator must be implemented to encrypt any data transferred between the SoC and any base station. The current design of the SoC includes a 32KHz clock source. However, a clocking unit capable of providing a wider range of operating frequencies will enable the SoC to adapt its frequency according to the application requirements. Finally, to enable the adoption of the SoC in a network architecture, a networking protocol must be implemented to handle reliable

communication in a multi-node system.

Appendix A

Publications

A.1 Completed

[FBY1] **Yahya, F.B.**; Patel,H. N.; Chandra, V.; Calhoun, B. H., "Combined SRAM Read/Write Assist Techniques for Near/Sub-Threshold Voltage Operation", *Design Automation Conference (DAC)* , 7-11 June 2015 (accepted as work-in-progress)

[FBY2] **Yahya, F.B.**; Patel,H. N.; Chandra, V.; Calhoun, B. H., "Combined SRAM Read/Write Assist Techniques for Near/Sub-Threshold Voltage Operation", *Quality Electronic Design (asQED), 2015 9th Asian Symposium on* , 4-5 August 2015

[FBY3] **Yahya, F.B.**; Mansour, M.M.; Tschanz, J.; Khellah, M.M., "Designing low-VTh STT-RAM for write energy reduction in scaled technologies," *Quality Electronic Design (ISQED), 2015 16th International Symposium on* , vol., no., pp.5,9, 2-4 March 2015

[FBY4] **Yahya, F.B.**; Patel, H.N.; Boley, J.; Banerjee, A.; Calhoun, B.H., "A Sub-threshold 8T SRAM Macro with 12.29nW Standby Power and 6.24pJ/access for Battery-less SoCs ", *Journal of Low Power Electronics and Applications*, vol. 6, no. 2, p. 8, May 2016.

[FBY5] Patel, H.N.; **Yahya, F.B.**; Calhoun, B.H., "Improving Reliability and Energy Requirements of Memory in Body Sensor Networks ", *2016 29th International Conference on VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID)*, Kolkata, 2016, pp. 561-562.

[FBY6] Roy, A.; Klinefelter, A.; **Yahya, F.B.**; et al., "A 6.45W Self-Powered IoT SoC with Integrated Energy-Harvesting Power Management and ULP Asymmetric Radios for Portable Biomedical Systems" *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 6, pp. 862-874, Dec. 2015.

[FBY7] Patel, H.N.; **Yahya, F.B.**; Calhoun, B.H., "Optimizing SRAM Bitcell Reliability and Energy for IoT Applications," *2016 17th International Symposium on Quality Electronic Design (ISQED)*, Santa Clara, CA, 2016, pp. 12-17.

[FBY8] Patel, H.N.; Roy, A.; **Yahya, F.B.** et al., "A 55nm Ultra Low Leakage Deeply Depleted Channel technology optimized for energy minimization in subthreshold SRAM and logic," *2016 46th European Solid-State Device Research Conference (ESSDERC)*, Lausanne, 2016, pp. 37-40.

A.2 Planned

[FBY9] **Yahya, F.B.**; et al. "A Battery-less 507nW SoC with Integrated Platform Power Manager and SiP Interfaces", submitted to VLSI 2017.

[FBY10] Patel, H.N.; Roy, A.; **Yahya, F.B.** et al., "A 55nm Ultra Low Leakage Deeply Depleted Channel Technology Optimized for Low-Power Internet of Everything", submitted to T-ED 2017.

[FBY11] Paper on BLE transmitter with University of Michigan titled "A 724W BLE Compatible FSK Transmitter for Wireless Beacon Applications".

[FBY12] Journal expansion of battery-less SoC paper - planned for JSSC.

[FBY13] Paper on low power FeRAM - planned for ESSCIRC 2017 (due date April 2017).

[FBY14] Paper on ultra low power voice activity detection.

[FBY15] Paper on Integration of SoC, FeAR and BLE into one package.

[FBY16] Paper on flexible on-chip bus enabling GALS and DVFS in self-powered Systems - planned for ESSCIRC 2017 (due date April 2017).

Appendix B

Acronyms

1T1R one transistor one resistor

1T1C one transistor one capacitor

2C two capacitor

2C2T two capacitor two transistor

2T two transistor

6T six transistor

8T eight transistor

ADC analog-to-digital-converter

AFE analog front end

AFIB atrial fibrillation

ALU arithmetic logic unit

AP anti-parallel (spin-torque)

BCU boost control unit

BL bit-line

BLB bit-line bar

BWL boosting the word-line

CBB cold-boot bus

CBC cold-boot controller

CBMS cold-boot management system

CMOS complementary metal oxide semiconductor

COTS commercial off-the-shelf

CU control unit

D2D die-to-die

DMA direct memory access

DMEM data memory

DMU data management unit

DVFS dynamic voltage and frequency scaling

ECG electrocadiography

EH-PPM energy harvesting platform power manager

FeAR ferroelectric-based auto-recovery system

Fe-RAM ferroelectric random access memory

Fe-FIFO ferroelectric first in first out

Fe-PROM ferroelectric programmable read only memory

FF fast NMOS fast PMOS corner

FIFO first in first out

FIR finite impulse response

FL free-layer

FS fast NMOS fast PMOS corner

FSK frequency shift keying

FW false write

GPP general purpose processor

GPIO general purpose input output

HS half-select (SRAM)

HSNM hold static noise margin

I2C inter integrated circuit

I_{C+} antiparallel-to-parallel switching threshold (spin-torque)

I_{C-} parallel-to-antiparallel switching threshold (spin-torque)

IMEM instruction memory

I_{rd} read current (spin-torque)

IoT internet of things

LCV_{DD} lowered-column supply

LPC low power controller

MAC multiply accumulate

MgO magnesium oxide

MTJ magnetic tunnel junction

NegBL negative bit-line

P parallel (spin-torque)

PDL pull-down left

PDR pull-down right

PGL pass-gate left

PGR pass-gate right

PL pinned layer (spin-torque)

PL plate line (ferroelectric)

PLL phase-locked loop

PM power monitor

PMA perpendicular magnetic anisotropy

POR power on reset

PUL pull-up left

PUR pull-up right

PV photovoltaic

RAM random access memory

R_{AP} anti-parallel resistance (spin-torque)

RBL read bit-line

RBM read burst mode

RBW read-before write

RDx row driver of row x

ROM read-only memory

RR heart rate interval between two R peaks

RSNM read static noise margin

R_P parallel resistance (spin-torque)

RV_{DD} raising the supply voltage

RWL read word-line

SAR successive approximation

SF source follower (spin-torque)

SF slow NMOS fast PMOS corner

SL source line (spin-torque)

SiP system-in-package

SoC system-on-chip

SNM static noise margin

SOR system power on reset

SPI serial peripheral interface

SRAM static random access memory

SS slow NMOS slow PMOS corner

STE short time energy

STT-RAM spin-torque transfer RAM

TEG thermoelectric generator

TT typical NMOS typical PMOS corner

UART Universal asynchronous receiver/transmitter

UDWL under-driven word-line

ULP ultra-low power

ULP-BUS ultra-low power bus

VAD voice activity detector

V_{BD} breakdown voltage of MTJ (spin-torque)

V_{CAP} super-capacitor voltage

VCO voltage controlled oscillator

V_{DD} supply voltage

V_{MAX} maximum voltage of technology

V_{MIN} minimum operating voltage

V_{MTJ} voltage across MTJ (spin-torque)

V_{REF} reference voltage

V_{rd} read voltage (spin-torque)

V_T transistor threshold voltage

V_{SS} ground voltage

VVSS virtual ground voltage

WID within die

WL word-line

WM write margin

WWL write word-line

XOR crystal power on reset

ZC zero crossing

Bibliography

[1] RJM Vullers, Rob van Schaijk, Inge Doms, Chris Van Hoof, and R Mertens. Micropower energy harvesting. *Solid-State Electronics*, 53(7):684–693, 2009.

[2] Yanqing Zhang, Fan Zhang, Y. Shakhsheer, J.D. Silver, A. Klinefelter, M. Nagaraju, J. Boley, J. Pandey, A. Shrivastava, E.J. Carlson, A. Wood, B.H. Calhoun, and B.P. Otis. A batteryless 19 u W MICS/ISM-band energy harvesting body sensor node SoC for ExG applications. *Solid-State Circuits, IEEE Journal of*, 48(1):199–213, Jan 2013.

[3] A. Klinefelter, N.E. Roberts, Y. Shakhsheer, P. Gonzalez, A. Shrivastava, A. Roy, K. Craig, M. Faisal, J. Boley, Seunghyun Oh, Yanqing Zhang, D. Akella, D.D. Wentzloff, and B.H. Calhoun. A 6.45 u W self-powered IoT SoC with integrated energy-harvesting power management and ULP asymmetric radios. In *Solid- State Circuits Conference - (ISSCC), 2015 IEEE International*, pages 1–3, Feb 2015.

[4] W. Lim, I. Lee, D. Sylvester, and D. Blaauw. 8.2 batteryless sub-nw cortex-m0+ processor with dynamic leakage-suppression logic. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, Feb 2015.

[5] B.H. Calhoun, A. Wang, and A. Chandrakasan. Modeling and sizing for minimum energy operation in subthreshold circuits. *Solid-State Circuits, IEEE Journal of*, 40(9):1778–1786, Sept 2005.

[6] B. H. Calhoun and A. Chandrakasan. Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90nm cmos. In *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.*, pages 300–599 Vol. 1, Feb 2005.

[7] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand. Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits. *Proceedings of the IEEE*, 91(2):305–327, Feb 2003.

[8] W. Lim, I. Lee, D. Sylvester, and D. Blaauw. 8.2 batteryless sub-nw cortex-m0+ processor with dynamic leakage-suppression logic. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, Feb 2015.

[9] N. Lotze and Y. Manoli. A 62mv 0.13 um cmos standard-cell-based design technique using schmitt-trigger logic. In *2011 IEEE International Solid-State Circuits Conference*, pages 340–342, Feb 2011.

[10] E. Seevinck, F.J. List, and J. Lohstroh. Static-noise margin analysis of MOS SRAM cells. *Solid-State Circuits, IEEE Journal of*, 22(5):748–754, Oct 1987.

[11] Zheng Guo, A. Carlson, Liang-Teck Pang, K. Duong, Tsu-Jae King Liu, and B. Nikolic. Large-scale read/write margin measurement in 45nm CMOS SRAM arrays. In *VLSI Circuits, 2008 IEEE Symposium on*, pages 42–43, June 2008.

[12] N. Verma and A.P. Chandrakasan. A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy. *Solid-State Circuits, IEEE Journal of*, 43(1):141–149, Jan 2008.

[13] S. Lutkemeier, T. Jungeblut, H.K.O. Berge, S. Aunet, M. Porrmann, and U. Ruckert. A 65 nm 32 b subthreshold processor with 9T multi-Vt SRAM and adaptive supply voltage control. *Solid-State Circuits, IEEE Journal of*, 48(1):8–19, Jan 2013.

[14] P. Meinerzhagen, O. Andersson, B. Mohammadi, Y. Sherazi, A. Burg, and J.N. Rodrigues. A 500 fW/bit 14 fJ/bit-access 4kb standard-cell based sub-VT memory in 65nm CMOS. In *ESSCIRC (ESSCIRC), 2012 Proceedings of the*, pages 321–324, Sept 2012.

[15] J. Kulkarni, M. Khellah, J. Tschanz, B. Geuskens, R. Jain, S. Kim, and V. De. Dual-VCC 8T-bitcell sram array in 22nm tri-gate CMOS for energy-efficient operation across wide dynamic voltage range. In *VLSI Circuits (VLSIC), 2013 Symposium on*, pages C126–C127, June 2013.

[16] Ik Joon Chang, Jae-Joon Kim, Sang Phill Park, and K. Roy. A 32 kb 10T subthreshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS. *Solid-State Circuits, IEEE Journal of*, 44(2):650–658, Feb 2009.

[17] Tae-Hyoung Kim, J. Liu, J. Keane, and C.H. Kim. A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme. In *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pages 330–606, Feb 2007.

[18] E. Karl, Zheng Guo, Yong-Gee Ng, J. Keane, U. Bhattacharya, and K. Zhang. The impact of assist-circuit design for 22nm sram and beyond. In *Electron Devices Meeting (IEDM), 2012 IEEE International*, pages 25.1.1–24.1.4, Dec 2012.

[19] Taejoong Song, Woojin Rim, Jonghoon Jung, Giyong Yang, Jaeho Park, Sunghyun Park, Kang-Hyun Baek, Sanghoon Baek, Sang-Kyu Oh, Jinsuk Jung, Sungbong Kim, Gyuhong Kim, Jintae Kim, Youngkeun Lee, Kee Sup Kim, Sang-Pil Sim, Jong Shik Yoon, and Kyu-Myung Choi. 13.2 a 14nm FinFET 128Mb 6T SRAM with VMIN-enhancement techniques for low-power applications. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pages 232–233, Feb 2014.

[20] C. Augustine, N. Mojumder, Xuanyao Fong, H. Choday, Sang Phill Park, and K. Roy. STT-MRAMs for future universal memories: Perspective and prospective. In *Microelectronics (MIEL), 2012 28th International Conference on*, pages 349–355, May 2012.

[21] T. Kawahara, R. Takemura, H. Takahashi, and H. Ohno. SPRAM (SPin-transfer torque RAM) design and its impact on digital systems. In *Electronics, Circuits and Systems, 2007. ICECS 2007. 14th IEEE International Conference on*, pages 1011–1014, Dec 2007.

[22] T. Andre, S.M. Alam, D. Gogl, C.K. Subramanian, H. Lin, W. Meadows, X. Zhang, N.D. Rizzo, J. Janesky, D. Houssameddine, and J.M. Slaughter. ST-MRAM fundamentals, challenges, and applications. In *Custom Integrated Circuits Conference (CICC), 2013 IEEE*, pages 1–8, Sept 2013.

[23] Dongsoo Lee, Sumeet Kumar Gupta, and Kaushik Roy. High-performance low-energy STT MRAM based on balanced write scheme. In *Proceedings of the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design*, ISLPED '12, pages 9–14, New York, NY, USA, 2012. ACM.

[24] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. Energy reduction for STT-RAM using early write termination. In *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, pages 264–268, Nov 2009.

[25] A. Driskill-Smith, D. Apalkov, V. Nikitin, X. Tang, S. Watts, D. Lottis, K. Moon, A. Khvalkovskiy, R. Kawakami, X. Luo, A. Ong, E. Chen, and M. Krounbi. Latest advances and roadmap for in-plane and perpendicular STT-RAM. In *Memory Workshop (IMW), 2011 3rd IEEE International*, pages 1–3, May 2011.

[26] Y.J. Lee, G. Jan, Y.J. Wang, K. Pi, T. Zhong, R.Y. Tong, V. Lam, J. Teng, K. Huang, R.R. He, S. Le, T. Torng, J. DeBrosse, T. Maffitt, C. Long, W.J. Gallagher, and P.K. Wang. Demonstration of chip level writability, endurance and data retention of an entire 8Mb STT-MRAM array. In *VLSI Technology, Systems, and Applications (VLSI-TSA), 2013 International Symposium on*, pages 1–2, April 2013.

[27] A. Sheikholeslami and P.G. Gulak. A survey of circuit innovations in ferroelectric random-access memories. *Proceedings of the IEEE*, 88(5):667–689, May 2000.

[28] M. Qazi, M. Clinton, S. Bartling, and A. P. Chandrakasan. A low-voltage 1 mb fram in 0.13 um cmos featuring time-to-digital sensing for expanded operating margin. *IEEE Journal of Solid-State Circuits*, 47(1):141–150, Jan 2012.

[29] J. T. Evans and R. Womack. An experimental 512-bit nonvolatile memory with ferroelectric storage cell. *IEEE Journal of Solid-State Circuits*, 23(5):1171–1175, Oct 1988.

[30] S. Khanna, S. C. Bartling, M. Clinton, S. Summerfelt, J. A. Rodriguez, and H. P. McAdams. An fram-based nonvolatile logic mcu soc exhibiting 100retention at vdd= 0 v achieving zero leakage with ¡ 400-ns wakeup time for ulp applications. *IEEE Journal of Solid-State Circuits*, 49(1):95–106, Jan 2014.

[31] V. Chandra, C. Pietrzyk, and R. Aitken. On the efficacy of write-assist techniques in low voltage nanoscale SRAMs. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2010*, pages 345–350, March 2010.

[32] Randy W. Mann, Jiajing Wang, Satyanand Nalam, Sudhanshu Khanna, Geordie Braceras, Harold Pilo, and Benton H. Calhoun. Impact of circuit assist methods on margin and performance in 6T SRAM. *Solid-State Electronics*, 54(11):1398 – 1407, 2010.

[33] James Boley, Vikas Chandra, Robert Aitken, and Benton Calhoun. Leveraging sensitivity analysis for fast, accurate estimation of SRAM dynamic write VMIN. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, pages 1819–1824, March 2013.

[34] M. E. Sinangil and A. P. Chandrakasan. Application-specific sram design using output prediction to reduce bit-line switching activity and statistically gated sense amplifiers for up to 1.9x lower energy/access. *IEEE Journal of Solid-State Circuits*, 49(1):107–117, Jan 2014.

[35] M. R. Stan and W. P. Burleson. Bus-invert coding for low-power i/o. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 3(1):49–58, March 1995.

[36] Daeyeon Kim, G. Chen, M. Fojtik, Mingoo Seok, D. Blaauw, and D. Sylvester. A 1.85fW/bit ultra low leakage 10T SRAM with speed compensation scheme. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pages 69–72, May 2011.

[37] Yih Wang, Hong Jo Ahn, U. Bhattacharya, Zhanping Chen, T. Coan, F. Hamzaoglu, W.M. Hafez, Chia-Hong Jan, P. Kolar, S.H. Kulkarni, Jie-Feng Lin, Yong-Gee Ng, I. Post, Liqiong Wei, Ying Zhang, K. Zhang, and M. Bohr. A 1.1 GHz 12 uA/Mb-leakage SRAM design in 65 nm ultra-low-power CMOS technology with integrated leakage reduction for mobile applications. *Solid-State Circuits, IEEE Journal of*, 43(1):172–179, Jan 2008.

[38] M.E. Sinangil, N. Verma, and A.P. Chandrakasan. A reconfigurable 65nm SRAM achieving voltage scalability from 0.25-1.2V and performance scalability from 20kHz-200MHz. In *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, pages 282–285, Sept 2008.

[39] M. F. Chang, M. P. Chen, L. F. Chen, S. M. Yang, Y. J. Kuo, J. J. Wu, H. Y. Su, Y. H. Chu, W. C. Wu, T. Y. Yang, and H. Yamauchi. A sub-0.3 v area-efficient l-shaped 7t SRAM with read bitline swing expansion schemes based on boosted read-bitline, asymmetric-v_{th} read-port, and offset cell VDD biasing techniques. *IEEE Journal of Solid-State Circuits*, 48(10):2558–2569, Oct 2013.

[40] J. J. Wu, Y. H. Chen, M. F. Chang, P. W. Chou, C. Y. Chen, H. J. Liao, M. B. Chen, Y. H. Chu, W. C. Wu, and H. Yamauchi. A large $\sigma v_{\text{th}}/vdd$ tolerant zigzag 8t SRAM with area-efficient decoupled differential sensing and fast write-back scheme. *IEEE Journal of Solid-State Circuits*, 46(4):815–827, April 2011.

[41] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[42] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. Tschanz, and V. De. A 2.3 nj/frame voice activity detector-based audio front-end for context-aware system-on-chip applications in 32-nm cmos. *IEEE Journal of Solid-State Circuits*, 48(8):1963–1969, Aug 2013.

[43] A. Shrivastava, D. Wentzloff, and B. H. Calhoun. A 10mv-input boost converter with inductor peak current control and zero detection for thermoelectric energy harvesting. In *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference*, pages 1–4, Sept 2014.

[44] Hector Ivan Oporta. An Ultra-low Power Frequency Reference for Timekeeping Applications. Master's thesis, Oregon State University, Oregon, USA, 2008.

[45] J. Pan and W. J. Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, March 1985.

[46] Doug E Lake and J. Randall Moorman. Accurate estimation of entropy in very short physiological time series: The problem of atrial fibrillation detection in implanted ventricular devices. *American Journal of Physiology - Heart and Circulatory Physiology*, 2010.

[47] Analog Devices. *ADXL362: Micropower, 3-Axis, 2 g/4 g/8 g Digital Output MEMS Accelerometer*, 2012. Rev. F.

[48] J. Myers, A. Savanth, D. Howard, R. Gaddh, P. Prabhat, and D. Flynn. 8.1 an 80nw retention 11.7pj/cycle active subthreshold arm cortex-m0+ subsystem in 65nm cmos for wsn applications. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, Feb 2015.

[49] D. Jeon, Y. P. Chen, Y. Lee, Y. Kim, Z. Foo, G. Kruger, H. Oral, O. Berenfeld, Z. Zhang, D. Blaauw, and D. Sylvester. 24.3 an implantable 64nw ecg-monitoring mixed-signal soc for arrhythmia diagnosis. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 416–417, Feb 2014.

[50] S. Emrani and H. Krim. Wheeze detection and location using spectro-temporal analysis of lung sounds. In *2013 29th Southern Biomedical Engineering Conference*, pages 37–38, May 2013.

[51] I. Lee, Y. Lee, D. Sylvester, and D. Blaauw. Battery voltage supervisors for miniature iot systems. *IEEE Journal of Solid-State Circuits*, 51(11):2743–2756, Nov 2016.

[52] A. Wang, A.P. Chandrakasan, and S.V. Kosonocky. Optimal supply and threshold scaling for subthreshold cmos circuits. In *VLSI, 2002. Proceedings. IEEE Computer Society Annual Symposium on*, pages 5–9, 2002.

[53] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70:614–636, 1996.

[54] P. Harpe, Hao Gao, R. van Dommele, E. Cantatore, and A. van Roermund. 21.2 a 3nw signal-acquisition ic integrating an amplifier with 2.1 nef and a 1.5fj/conv-step adc. In *Solid- State Circuits Conference - (ISSCC), 2015 IEEE International*, pages 1–3, Feb 2015.