Integrated Circuit Design for Miniaturized, Trackable, Ultrasound Based Biomedical Implants

Yihan Zhang

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020
Integrated Circuit Design for Miniaturized, Trackable, Ultrasound Based Biomedical Implants

Yihan Zhang

This thesis focuses on the design of an ultrasonography compatible implantable sensor platform, as a novel approach that implements a miniaturized, battery-less, real-time trackable parallel biosensing system. In addition to the frontend circuit, a sub-nW fully integrated pH sensor is designed in a way that can be easily integrated with the proposed sonography-compatible sensor platform. Combining the two integrated circuits together, the whole system will be able to map in vivo physiological information acquired from a distributed set of sensors on top of the ultrasound movie, leading to the idea envisioned as “augmented ultrasonography”.

Implemented in a 0.18 µm technology, an ultrasound power and data frontend circuit is designed to enable medical sensing implants to operate in an ultrasonography compatible way. When placed within the field of view of an imaging transducer, the frontend circuit harvests the power through a piece of piezo crystal from a minimally modified brightness-mode (B-mode) ultrasound imaging process that is commonly adopted in modern medical practices. The implant can also establish bi-directional data communication channels with the imaging transducer, allowing data to be transmitted in a way synchronized to the frame rate of the B-mode film. The design of the circuit is made possible by a combination of ultra-low-power circuit techniques and novel frontend circuit topologies, as imaging ultrasound waves in the form of short pulses with extremely low duty cycle poses challenges that has not previously seen in other implantable sensor systems. The proposed prototype achieves a total area of 0.6 mm$^2$ for the integrated circuit (IC), as well as 71 mm theoretical maximum implantable depth (up to 40 mm is verified experimentally). These two together give opportunities for this design to become the next generation solution for deep-tissue bio-sensing.
Realized using the same 0.18 µm technology, the fully integrated pH sensor is designed to deliver accurate pH readouts, at a reasonable speed of 1 sample per second, while consuming only 0.72 nW of power. Using an ion-sensitive field effect transistor (ISFET) and reference field effect transistor pair (REFET), the IC requires minimum additional post fabrication to deliver 10-bit resolution pH readouts at an end-to-end sensitivity of 65.8 LSB/pH. When working as a standalone device, this work advances the state-of-the-art of ISFET based pH sensor design. With an addition of 0.46 mm² of area, it is possible to integrate it with the ultrasound sonography compatible implant platform. This potential integration will further advance the vision of the augmented ultrasonography: real-time display of physiological information in a B-mode film, with the help from a distributed bio-sensor system for deep-tissue physiology monitoring.
# Table of Contents

List of Tables .................................................................................................................. v

List of Figures .................................................................................................................. vi

Acknowledgments ............................................................................................................ x

Chapter 1: Introduction ...................................................................................................... 1
  1.1 Implantable Biomedical Sensors ............................................................................. 3
  1.2 Deep Tissue Implant .............................................................................................. 4
  1.3 Thesis Outline ......................................................................................................... 5

Chapter 2: Ultra-Low-Power Integrated Circuit Design ..................................................... 7
  2.1 MOSFET in Deep-Subthreshold Operation .............................................................. 8
     2.1.1 MOS Modeling in Deep-Subthreshold .............................................................. 9
     2.1.2 Simulating MOSFETs in Deep-Subthreshold ................................................ 13
  2.2 Deep-subthreshold Analog Design ........................................................................ 13
     2.2.1 Transconductance Efficiency and Intrinsic Gain .......................................... 14
     2.2.2 Speed Consideration ..................................................................................... 16
     2.2.3 Noise Consideration ..................................................................................... 17
     2.2.4 Mismatch Consideration .............................................................................. 18
2.2.5 Operational Transconductance Amplifier Design Flow ............... 22
2.2.6 Summary and Discussion .................................................. 23
2.3 Low-Voltage Digital Design .................................................. 24
  2.3.1 Complementary CMOS in Deep-Subthreshold ....................... 25
  2.3.2 Dynamic Leakage Suppression Logic .................................... 28
2.4 Building Blocks for Sub-nW Systems ........................................ 30
  2.4.1 Standard Cell Library ..................................................... 30
  2.4.2 Low Leakage ESD ......................................................... 32
2.5 Summary ................................................................. 35

Chapter 3: An Ultrasonography Compatible Implant ............................. 38
  3.1 Background ................................................................. 39
    3.1.1 Power Delivery for Deep Tissue Implants ......................... 39
    3.1.2 Ultrasonography ....................................................... 41
  3.2 System Level Design ........................................................ 44
    3.2.1 Sonography Ultrasound Wave Seen from the Implant ............. 45
    3.2.2 Top-Level Modular Design ............................................. 47
    3.2.3 Modification to the Existing Imaging Process .................... 49
    3.2.4 Ultrasound System .................................................... 50
  3.3 Circuit Design ............................................................ 50
    3.3.1 Switch-Only Rectifier .................................................. 51
    3.3.2 Voltage Regulator ..................................................... 56
    3.3.3 Clock Recovery .......................................................... 60
<table>
<thead>
<tr>
<th>Section Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.4 Downlink Data Recovery</td>
<td>61</td>
</tr>
<tr>
<td>3.3.5 Uplink Data Modulator</td>
<td>63</td>
</tr>
<tr>
<td>3.3.6 Periphery Circuit Blocks</td>
<td>67</td>
</tr>
<tr>
<td>3.3.7 Link Layer and Application Layer</td>
<td>70</td>
</tr>
<tr>
<td>3.4 Performance Verification</td>
<td>71</td>
</tr>
<tr>
<td>3.4.1 Post Fabrication</td>
<td>71</td>
</tr>
<tr>
<td>3.4.2 Electrical Measurement</td>
<td>73</td>
</tr>
<tr>
<td>3.4.3 Ultrasound Based Characterization</td>
<td>76</td>
</tr>
<tr>
<td>3.4.4 Multi-Implant Operation under B-Mode Sonography</td>
<td>82</td>
</tr>
<tr>
<td>3.5 Summary</td>
<td>82</td>
</tr>
<tr>
<td>Chapter 4: A Sub-nW Integrated pH Sensor</td>
<td>86</td>
</tr>
<tr>
<td>4.1 Background</td>
<td>86</td>
</tr>
<tr>
<td>4.2 System Design</td>
<td>87</td>
</tr>
<tr>
<td>4.2.1 Top Level Design</td>
<td>87</td>
</tr>
<tr>
<td>4.2.2 ISFET and REFET</td>
<td>89</td>
</tr>
<tr>
<td>4.2.3 ISFET Frontend</td>
<td>93</td>
</tr>
<tr>
<td>4.2.4 ADC</td>
<td>98</td>
</tr>
<tr>
<td>4.2.5 Switched-Capacitor Amplifier</td>
<td>101</td>
</tr>
<tr>
<td>4.2.6 Periphery Circuit Blocks</td>
<td>104</td>
</tr>
<tr>
<td>4.2.7 Digital Interface</td>
<td>106</td>
</tr>
<tr>
<td>4.3 Performance Verification</td>
<td>108</td>
</tr>
<tr>
<td>4.3.1 Post Fabrication and Packaging</td>
<td>109</td>
</tr>
</tbody>
</table>
4.3.2 PCB Design .................................................. 112
4.3.3 FPGA and Software Design .............................. 114
4.3.4 pH Measurement ........................................... 116
4.3.5 Electrical Measurement ................................. 121
4.4 Summary ....................................................... 122

Chapter 5: Conclusion ............................................. 125
  5.1 Summary of Contribution .................................. 125
  5.2 Future Work ................................................ 126
    5.2.1 The Ultrasonography Compatible Sensor Frontend .. 126
    5.2.2 The Sub-nW Integrated pH Sensor ..................... 127
  5.3 Final Remarks .............................................. 128

References ...................................................... 129
List of Tables

2.1 Available cells in the standard cell library (O3V). ........................................ 32

3.1 Frame structure for link layer. ................................................................. 70

3.2 Application layer instructions. ................................................................. 70

3.3 Comparison between the proposed tracking mechanism and RSSI. ............... 83

3.4 Comparison between this work and other ultrasound based sensor platforms. .... 84

4.1 Desired performance of the designed pH sensor ....................................... 89

4.2 Frame structure for the output serial data in the standalone implementation. .... 107

4.3 Register mapping for the pH sensor control signals. ................................. 107

4.4 Comparison between this work and other digital pH sensor systems, or low power pH sensor systems. ................................................................. 124
List of Figures

2.1 Measurement setup for deep-subthreshold drain current. (A) Schematic, and (B) photo from probe station. ...................................................... 11

2.2 A measured result of deep-subthreshold current compared to simulation data. . . . . 12

2.3 Simulated transconductance efficiency $g_m/I_D$ and intrinsic gain $g_m/g_{DS}$ for a thick oxide 3.3 V NMOSFET. ................................................................. 15

2.4 Schematic of a CMOS complementary inverter. ........................................ 26

2.5 Schematic of a DLS inverter, and its simulated time domain waveform under Fo4 configuration. ................................................................. 28

2.6 The layout of an inverter and a NAND gate in the standard cell library. .......... 31

2.7 Schematic of the proposed ultra-low-leakage power clamp. ...................... 33

2.8 The simulation testbench setup for the proposed ESD structure under HBM ESD events. The anticipated discharge path is highlighted with a red arrow. ........ 34

2.9 Simulated transient waveform on $V_{DD}$ upon an 1 kV HBM ESD event. ....... 35

2.10 Comparison of the leakage current between the proposed power clamp in simulation as well as in measurement, and that to a commercially available ESD structure. . . 36

3.1 Three different cases for spatial power distribution when powering up deep tissue power harvesting implants. ...................................................... 40

3.2 The working principle of pulse-echo mode ultrasound imaging. ................ 42

3.3 The working principle of pulse-echo mode ultrasound imaging. ................ 43

3.4 Concept illustration of the ultrasound-based biomedical implant designed to work under medical sonography ............................................. 45
3.5 Ultrasound waveform received at the implant as a linear array transducer scans across the body. ................................................. 47
3.6 Block diagram for the sonography compatible implant. ............... 48
3.7 Measured impedance of different piezo crystals under different conditions. .......... 52
3.8 Schematic of the switch-only rectifier. .................................. 53
3.9 Conceptual waveforms for the switch only rectifier. .................. 54
3.10 Schematic of the comparator used in the active diode. ................. 55
3.11 Schematic of the voltage reference. ..................................... 57
3.12 Schematic of the voltage regulator. ...................................... 59
3.13 Schematic and waveform of the clock recovery circuit. ............. 60
3.14 Schematic and waveform of the downlink data recovery circuit. .... 62
3.15 Schematic of the uplink data modulator. ................................ 64
3.16 Illustration of the case when uplink corrupts the clock recovery without a proper modulator. ......................................................... 65
3.17 Schematic of the 5-bit tunable falling edge delay element. .......... 66
3.18 Waveforms of the key signals in the uplink data modulator. ......... 67
3.19 Schematic for the power-on-reset. ....................................... 68
3.20 Schematic for one bit fuse based ID generator. ......................... 69
3.21 Die photo for the designed IC as the core of an ultrasound sonography compatible implant. .................................................. 72
3.22 Local die photo before and after the etching of the fuse. ............. 73
3.23 A photo of the fully packaged ultrasound sonography compatible implant. ...... 74
3.24 The abstract illustration of the electrical testing setup. ............... 74
3.25 DC power regulation characteristic of the voltage regulator measured using probe station. ..................................................... 75
3.26 Rectifier power harvesting efficiency measured across different duty cycle ultrasound pulses. ................................................................. 76
3.27 Time domain waveform of the high power delivery from the IC, at its output node. ................................................................. 77
3.28 Measured ultrasound pulses from the imaging transducer, captured by a hydrophone. 78
3.29 Time domain waveforms for the unregulated $V_{CC}$ and regulated $V_{DD}$. ................................................................. 79
3.30 Waveforms for clock and downlink data recovered on the chip, and the uplink data from the chip under imaging mode ultrasound. ................................................................. 80
3.31 Recorded ultrasound waveforms at the transducer, when the uplink data is an “1” (red) or an “0” (blue). ................................................................. 80
3.32 Pressure at a specific location in the reconstructed image, where an implant is found, plotted against the number frames during acquisition. ................................................................. 81
3.33 Example of an acquired B-mode image of a device, and its $E_b/N_0$’s spatial distribution. 81
3.34 Minimum power required from the transducer, as well as the $E_b/N_0$ as a function of implant depth. ................................................................. 82
3.35 Experiment setup to demonstrate parallel operation of two devices under the FoV of one imaging transducer. ................................................................. 85

4.1 Top level conceptual block diagram. ................................................................. 88
4.2 An illustration of the sensitivity of the surface charge to the change in pH. SiO$_2$ passivation is used as an example. ................................................................. 91
4.3 A picture of the conceptual cross-section of the ISFET-REFET pair. ................................................................. 92
4.4 Topology of the traditional CVCC circuit. ................................................................. 95
4.5 Topology of the proposed pseudo-differential source follower pair. ................................................................. 96
4.6 Simulated distortion in a pseudo differential source follower pair. ................................................................. 98
4.7 Schematic of the 10-bit monolithic switching SAR ADC. ................................................................. 99
4.8 Cross-section view of the unit capacitor. ................................................................. 99
4.9 Schematic of the comparator used in the ADC. ................................................................. 100
4.10 Layout of the comparator used in the ADC. ........................................... 101
4.11 Schematic of the switched capacitor amplifier. ................................. 102
4.12 Schematic of the fully differential amplifier. .................................. 103
4.13 The common mode voltage generator. .............................................. 104
4.14 The digitally configured current source. ......................................... 105
4.15 Die photo for the standalone testable fully integrated pH sensor IC. .... 109
4.16 Illustration of the post fabrication flow. .......................................... 110
4.17 Illustration of several key steps during the encapsulation process. (1) application of a line of epoxy, (2) epoxy after curing, (3) the full inner dam after curing, and (4) the full encapsulation. .............................................. 112
4.18 Block diagram of the 3 boards designed for real time pH measurements. .... 113
4.19 Soldered PCB’s for the pH measurement system. ............................ 114
4.20 Soldered PCB’s for the pH measurement system when the boards are connected as designed. .................................................. 114
4.21 Block level diagram for the Verilog modules used in Opal Kelly XEM6010 FPGA. 116
4.22 Relations between different threads used in the software. .................... 117
4.23 The user interface of the software used for testing. ............................ 118
4.24 The time domain behavior of the pH sensor between pH=6.00 to pH=10.00. .... 119
4.25 End-to-end sensitivity of the pH sensor. ......................................... 120
4.26 5-hour long measurement demonstrating the stability of the sensor. ........ 120
4.27 Transient response and current consumption measured using probe station. ... 122
4.28 Performance summary under different clock frequency and bias current. .... 123
Acknowledgments

It has been indeed a long journey since I first came to the United States of America, and looking back now, my mental state has evolved quite a bit, thanks to the educational experience that I have the fortune to enjoy here as a member of the Bioelectronic Systems Lab at Columbia University in the City of New York.

I would like to first thank my advisor, Dr. Kenneth L. Shepard, who dedicates his life creating a free, inspiring, and multi-disciplinary research environment, that I believe is unique and outstanding in academia nowadays. Ken’s great passion towards research, creativity for novel approaches, and a broad spectrum of knowledge on many disciplines all inspire me as a novice engineer and scientific researcher.

I would also like to thank the members of my committee: Professor Yannis Tsividis, Professor Lars Dietrich, Professor Samuel Sia, and Professor Parag Chitnis for their time and commitment to the completeness of my Ph.D. program.

I would particularly like to thank Peijie Ong for his great support during the time when I lost almost all my confidence in research. Your mental support as well as the language editing service are the key reasons for my first paper to ever get accepted. Hope your cookie factory grows more prosperous than ever.

On top of that, I also want to thank Steven Warren, Sefi Vernick, Eyal Aklimi, and Daniel Bellin for the early time tutoring when I first joined the group. Only looking back now do I realize how terrible I was, not just limited to academic performance. And your patience and support were among the first things that I could rely on.

It is also a great fortune for to meet many kind and inspiring friends and coworkers during my seven years at Columbia University. This certainly includes, but not limited to Cheng Tan, Fengqi Zhang, Daniel Fleischer, Siddharth Shekar, Jeffery Elloian, Jeffery Sherman, Chen Shi, Tiago
Costa, Filipe Cardoso, Andreas Hartel, Girish Ramakrishnan, Rizwan Huq, Kukjoo Kim, Nanyu Zeng, Taesung Jung, Prashant Muthuraman, and everyone else in the Bioelectronics Systems Lab. I need to stop the list here before this takes over the whole acknowledgment page. And please accept my sincere apology if you don’t find your name explicitly listed above. If you read my thesis, I am more convinced than enough that you are my friend.

I would also like to thank William-Cole Cornell, Anastasia Bendebury, Hassan Sakhtah, Lisa Kahl, Jeanyoung Jo, Bryan Wang and Professor Lars Dietrich (again) from Dietrich Lab, for all the fun times we had together with Pseudomonas Aeruginosa biofilms.

There is also a couple of Professors that I would like to express my special gratitude to. Your appearance at a certain stage in my life has helped me significantly, possibly in a way you have not expected. This includes but not limited to: Professor Peter Kinget, who I find can best explain how circuit works within all the circuit classes that I have taken; Professor Hanh-Phuc Le, as you probably did not realize that your complement to my work is the very first opinion I have every received in a public professional presentation; Mr. Maysam Ghovanloo, your insightful analysis into my work has inspired me so much; Professor David Blaauw, a quick chat with you surprisingly inspired me to look at my life from an engineering perspective; Professor Eitan Grinspun, for a wonderful year and fantastic summer in the field of computer graphics; and finally, Professor Qiao Lin, the interaction with you has significantly broadened my understanding to humanity.

Last but not the least, I would like to thank my father and mother for bringing me to this troubling yet extremely wonderful world. I feel great to be alive, and grateful to be alive. My thanks also go to all of my family members. You have always been very helpful, and I hope I could make your life better too in the future. And finally, I would like to express my deepest gratitude to my wife Yi Wang, without whose support I would not be the person I am, and with whom I choose to spend the rest of my life together.
Chapter 1: Introduction

Since its invention in the 1950s, the design of a modern IC has become one of the most sophisticated engineering practice every known to human beings, within which billions of metal-oxide-silicon-field-effect-transistors (MOSFET) need to be scrambled into a millimeter scale tiny solid-state chip in an orderly way, with layers of metal traces woven on top, to perform tasks demanded by the market, and to bring the designer a decent income. This view we have today, was probably first clearly envisioned by Gordon Moore, in his famous Moore’s law \[1\]. Over the years, Dennards scaling \[2\] has been the highway towards Moores vision, yet it has become a bumpy one soon after the year 2000, when the characteristic feature of the MOSFET went into the region known as deep sub-micron. Since then, the dooms day tale of the end of Moore’s law started to pick up its heat. However, countless scientists and engineers around the world still dedicated their life to pave the road forward. As of time of this writing, devices with 7 nm feature size have already been commercialized, while that with 5 nm are expected to be open for commercial usage soon. But behind this, what is the force that has driven people to this point? What is the gold that people have been chasing for by keeping Moores law well held?

My answer to this question, after this lengthy journey as a graduate student, is the predictable profit along the curve Moore draws. With more devices compressed into a same silicon grain, the same functionality can be achieved at a lower cost, at a higher speed, and possibly also at a lower power consumption. The demand for more computational resources has been well boosted by advanced scientific work, interactive entertainment, personalized service, and omnipresent connectivity. And another step of device scaling simply guarantees the address for such a demand, while simultaneously advocates for more. Moore’s law has thus promised prosperity, not only to people working in the integrates circuit industry, but also to everyone whose life has already been strongly coupled to pieces of electronics.
Ages of fast expansion in the field has made integrated circuits more and more capable of processing electrical signals, yet physical laws have never failed to disillusion our hope for infinity. The current journey will come to an end, however, when we look around, modern electronics has grown so powerful, where an IC in the size of a human cell (about 100 µm in diameter) can already incorporate up to about a million of logic components. Is it possible, to look aside from the Moore’s direction to the utopia where computation has no cost at all, but to walk to a different way, known broadly as “more than Moore”, and use what we have achieved in the modern IC technology to address human being’s never ending desire for a longer and more pleasant life? That is the question which has driven the emergence of the field known as bioelectronics.

It is important to note that the exact definition of bioelectronics remains ambiguous. It can refer to different ideas in different contexts. Possibly the most conservative definition for this concept is given by the National Institute of Standard and Technology (NIST) [3], where bioelectronics is referred to as “the discipline resulting from the convergence of biology and electronics”. One category of ideas falls into this definition, where biological systems are engineered to behave like electronics using the modern day knowledge of signal and systems. This idea is probably better described in [4], in which bioelectronics was defined as “the use of biological materials and biological architectures for information processing systems and new devices”, leading to the vision of VLSI-like network composed of biological structures, like bi-molecular devices and biosensors. So far, works within this field suffer from limited progress, only with a few pioneering works like neural circuits [5]. But so far, most of the work along this line exists only in research labs, and there is no clear hint of a massively production in the industry. Another category of research work in bioelectronics focuses on the design of electronic systems that support biomedical diagnosis and treatments, which is the field this work belongs to. And more specifically, this thesis focuses on the design of a miniaturized, battery-free, power harvesting medical implant system for sustainable and affordable healthcare.
1.1 Implantable Biomedical Sensors

Many medical diagnostic techniques involve the removal of a part of the human body as sample, such as the person’s secretions, body fluids, or tissue with signs of lesions. This sample then undergoes careful examination that sometimes involves steps of processing. To map the biological sample to a finite set of known disorders, the information hidden within are extracted through biological processes, chemical reactions, physical experiments, and direct recognition from a trained personnel, before typed into an electronic device, where the information exists in its final form as a finite set of electrical signal.

The idea behind an implantable biomedical sensor, however, promises a different scenario. The entire sensor system is now directly placed within the body, around the area of interest. The desired biological information needs to be converted locally into electrical domain, a signal modality of which processing and long-distance transmission is the easiest thanks to well-developed IC technology. This vision immediately give rise to a set of desired properties for well-designed implantable biomedical sensors. Firstly, the acquisition of the biological information should not disturb the proper function of the body, which requires the sensor to be designed in a miniaturized way, with a proper, bio-compatible coating, remain wireless, and under a certain heat dissipation budget. Secondly, the measurement conducted by the sensor needs to deliver reliable and accurate results. Thirdly, the lifetime of the implants needs to be maximized, to avoid frequent surgery required to replace them.

Combining some of the desired properties helps us draw an illustration of what an ideal implant system will look like. One or more tiny grain of sensor that is bio-compatible are hiding underneath the skin. During operation, they send accurate biological information out of the body to another host device. This scientific fiction like image is certainly attractive, yet when an engineer takes a ruler out, and starts to quantify each of the desired qualities, things are then brought down to a struggle along a rocky path. How small? How bio-compatible? How accurate are the measurements? How fast is the data transmission? And if all above looks great, then how expensive it is? There are few
recent works that has delivered a satisfactory score to the questions above, yet the development is even more limited for implants that are designed to interact with the internal organ sitting deeply under the surface of the skin.

1.2 Deep Tissue Implant

Human body is not designed to be integrated with electronics. Quite a fraction of the body structure is designed to provide an isolated environment free from interference, electrical and mechanical, to safeguard the correct functionality of the core biological systems. As the location of the implants get closer to the organs, powering and data transmission now need to breach more layers of protection, viewed by the engineers as “a higher path loss” for both power and signal. On top of this, internal organs also induces unconscious and often constant movement, as a background fulfillment of key life supporting functionalities and a prevention to certain kinds of conscious suicide. If now a tiny implant is placed around such organs, its relative movement can easily exceed its feature size. To make the matter worse, the closest place to put a hosting device is now also further from the implant, as the implant goes deeper into the tissue. Finding and tracking such devices become an emerging challenge that has not been a problem significant enough for close-to-surface implant designers to notice in the past.

Deep tissue implant is the core topic that will be discussed in this thesis. To deliver desired features discussed previously, this work instead look at this problem at a different angle, and provides a unique and novel solution to the next generation deep-tissue implants. In the past several years, an improved understanding of deep-subthreshold circuit operation has significantly pushed the lower boundary of power consumption to implement functionalities that only needs performs at a low speed. This opens up the possibility to design implants that uses energy from techniques that are considered for imaging purposes only, where spatial power density is much lower than custom designed power delivery channels. Yet from decades of medical practice, these techniques have been well studied for their abilities to investigate organ behavior deep down the skin, as well as how to operate them safely. The imaging technique of interest within the scope of this thesis, is
the B-mode medical ultrasound imaging, or B-mode ultrasonography. Through a combination of ultra-low-power circuit design and B-mode imaging, this thesis will soon demonstrate, to the best knowledge of the author, the design of the world’s first miniaturized, trackable implant platform for the next generation deep tissue bio-sensing applications.

1.3 Thesis Outline

This thesis assumes the readers has a reasonable background in integrated circuit design and analysis. Subjects discussed in common undergraduate level textbooks, like [6] and [7] may be treated as known to the reader, and thus little explanation will be provided in this work. On top of that, the writing of this thesis is done in a much more honest way: both the advantage and disadvantage of certain design choices will be discussed, an approach that is sadly increasingly rare in modern day academic publications. With that, the author hopes the arrangement of this thesis will deliver a more comprehensive understanding towards the works that my life as a graduate student has dedicated to, including their novelty, the advance of the state of the art in performance, as well as their limitations, and problems needs to be addressed before commercialization become possible. All the IC based work presented in this thesis comes from two taped-out chip sets. They are code-named as USTAG V2.71 (July 18, 2018), and USTAG V2.7 (November 6, 2019). Ultimately, all the works presented here suffer from intrinsic trade-off between their quality and the time spent by the designer.

The rest of the thesis is arranged as follows. Chapter 2 provides a quick and rough walkthrough of deep-subthreshold analog and digital circuit designs. This technique is the backbone to implement the proposed sensor platform. Two important building blocks are introduced at the end of this chapter, marking the starting point of the authors original contribution to the field. These blocks are shared in the designs introduced later. Chapter 3 details the design of a power and data frontend circuit that can be used as an ultrasound B-mode imaging compatible biomedical sensor platform. Chapter 4 describes a sub-nW pH sensor that is designed to be integrated with the sensor platform for in vivo pH sensing, as a demonstration of the ultrasonography based frontend circuit’s potential.
Chapter 5 offers conclusion, and a brief glance on the potential improvements in the future to further extend the scope of this work.
Chapter 2: Ultra-Low-Power Integrated Circuit Design

Biological signals usually occupy the lower side of the frequency band in the spectrum. Probably the highest frequency signal is the fast-spiking behavior of the action potential in the neural system [8, 9, 10, 11], where up to 7 kHz of effective cutoff frequency has been found to be beneficial for recording purposes. Commonly measured electrophysiological signals, like electroencephalography (EEG), have been traditionally studied in a bandwidth below 13 Hz [12], with the sampling frequency going up to 1 kHz. Even modern IC designs for EEG sensing purposes [13, 14, 15] adopt sampling frequencies no higher than 125 Hz. Other than neural activity, most health-related physiological information varies at an even lower frequency, typically around or lower than 1 Hz, like body temperature, pH, lactate concentration, heart rate, blood pressure, etc.

Circuit design for biomedical applications can then take advantage of the slow-varying nature of biological signals. A reduction of operating bandwidth can be traded into other performance metrics. A lower power consumption is one of them. Implementing biomedical applications under extremely limited power budget is also one of the key challenges in the work presented in this thesis. The ultrasonography compatible implant system, which will be presented in Chapter 3, faces a limited amount of power available to harvest by its working principle. The integrated pH sensor, which will be presented in Chapter 4, utilizes the slow varying nature of pH in biological systems to drastically reduce its power consumption to a level that can be driven by the ultrasonography compatible frontend circuit.

In this chapter, general considerations are presented involving both transistor and topology level when designing sub-nW integrated systems. The principles and techniques discussed in this chapter apply to all designs documented in later chapters. Since a single transistor with a reasonable size (like 1 μm/1 μm), when conducting 1 nA, is already operating in deep-subthreshold region, the somehow ambiguous term of “ultra-low-power” is replaced with “deep-subthreshold” or “sub-nW”
in the technical context that is to follow whenever appropriate.

The rest of this chapter is organized as the following. First, a brief introduction of the deep-
subthreshold MOSFET behavior will be introduced, accompanied with the challenge it brings to
modeling, and thus the interpretation of the simulation results. Then, a brief walkthrough is offered
on the design of both digital and analog circuits using MOSFET biased in deep-subthreshold. After
that, the design of some basic and necessary circuit building blocks are introduced, including a
close-to-minimal standard cell library, and an electrostatic discharge (ESD) power clamp. These
designs are shared in later works. Finally, this chapter concludes with a brief discussion on the
limitation of the considerations presented here.

2.1 MOSFET in Deep-Subthreshold Operation

When biased under deep-subthreshold operation, the week inversion current of a MOSFET, as
a function of the voltage on its four terminals, can be written as \[ I_{DS} = \frac{W}{L} \mu \left( \frac{\sqrt{2q \varepsilon S N_A}}{2\sqrt{2\phi_F + V_{SB}}} \right) \phi_T^2 e^{(V_{GS} - V_{TH})/n\phi_T} (1 - e^{-V_{DS}/\phi_T}). \] (2.1)

We can further simplify this equation by defining:

\[ I_S = \mu \frac{\sqrt{2q \varepsilon S N_A}}{2\sqrt{2\phi_F + V_{SB}}} \phi_T^2. \] (2.2)

This term is a function of process parameters, temperature of operation, and body bias. Grouping
them together offers a simpler expression, while exposing most of the parameters that a designer
has control over. Thus the deep-subthreshold current equation can be re-written as:

\[ I_{DS} = I_S \frac{W}{L} e^{(V_{GS} - V_{TH})/n\phi_T} (1 - e^{-V_{DS}/\phi_T}), \] (2.3)
the form this thesis will stick to for the rest of the discussion for convenience. The controllable parameters are the geometry of the device and the terminal voltages. Although physically sound, this equation overlooks certain important mechanisms that interfere with the operation of a MOSFET, which has particular importance to deep-subthreshold operations. This will be discussed shortly next, when we consider the physical modeling of the transistor. Yet still, the equation presented above provides the most accurate quantitative understanding of subthreshold behavior of a MOSFET. Thus in practical design process, the result given from this equation is treated as the first order solution that defines the functionality of the circuit, and mechanisms that cannot be expressed in an analytical fashion well are treated as second order effects, which will be kept under control in a quantitative fashion.

2.1.1 MOS Modeling in Deep-Subthreshold

The powerful modern-day simulator has evolved into one of the dominant players in the choice of design decision, due to its superior capability in the prediction of complicated behaviors from MOSFETs that has hundreds of modeling parameters. However, such predictions are not as reliable in deep-subthreshold region, mostly because the modeling effort is not very well spent to optimize the accuracy in this region. It makes sense as the market share remains negligible for ICs depending heavily on deep-subthreshold behavior of MOSFETs. Therefore, it is the designer’s job to compensate for such modeling errors, making sure that they do not interfere with the functionality of the circuit, and the degradation of the performance from such errors are limited to an acceptable level.

The work presented in this thesis mainly uses 180-nm-family technologies. Devices available including core devices with different threshold voltages, and I/O devices that are approximately equivalent to those in a 350 nm node. In these technologies, gate leakage is luckily less of a concern, and can be further eliminated from design consideration by using I/O devices that has an even thicker gate oxide.

Gate induced drain leakage, or GIDL for short, is possibly the most notable physical effect that is
not well modeled. This effect can be understood intuitively in the following way. In an NMOSFET, when the gate is biased negatively, band bending and depletion will happen in the gate-overlapped n-doped region of the drain, if the drain node is biased positively. When the bending of the band exceeds that of the bandgap, direct tunneling will happen between drain and the p-doped bulk. A more detailed explanation of this phenomena and its modeling in Berkeley Short-channel IGFET Model version 4 (BSIM4) can be found in [17]. The most common place when GIDL plays an important role, is when “super-cutoff” happens: the situation where the source voltage is elevated above the ground to generate a negative $V_{GS}$ for subthreshold leakage reduction. If the gate is biased too negative, GIDL can take over subthreshold leakage, and become the dominant leakage mechanism.

To provide a quantitative understanding of this phenomenon, a single standard $V_{TH}$ core 1.8V NMOSFET with size of 50 mm/0.18 µm is taped-out in chipset USTAG-V2.7. A measurement of GIDL current is performed using the probe station setup shown in Figure 2.1. The measured result compared to simulation and theoretical fittings are plotted in Figure 2.2. The NMOSFET is protected with custom designed ESD structures that will be described later in Subsection 2.4.2. The ESD circuit gives rise to a negligible error in the measurement when the power clamp is properly biased, and this has been verified by cross checking the result with that measured on the same transistor protected by commercial ESD devices.

In Figure 2.2, different $V_{GS}$ regions are labeled using background colors: a red background indicates Equation 2.3 predicts a more accurate result than that of from the commercial model (based on BSIM 4.5), and blue shades the region where moderate and strong inversion starts, such that subthreshold equation predicts a worse result than model based simulation. It is worth noting that the $V_{SB}$ in both test setup and simulation is set to 800 mV, contributing to a non-negligible level of body effect. It is indeed a common and realistic use case of super-cutoff. In the cases where the source terminal is connected to the body of the transistor, a significant improvement of model accuracy can be found due to better modeling at bias points commonly seen, but theoretical calculations still predicts better than what the model can achieve. It is clear that in this figure, the
simulated GIDL effect is overshadowed by the inaccurately modeled subthreshold current, and an over-prediction of almost $100 \times (g_{\text{min}} \text{ set to 0 in this simulation})$ subthreshold current is found from simulation. With measurements like this, the optimal minimum leakage point can be found with high level of confidence, that leads optimal choices in design. In this particular case, $V_{GS} \approx -200 \text{ mV}$ is found as the optimal gate voltage for the minimum leakage current.

Another important mechanism is drain induced threshold shifting, or GITS \cite{17}. GITS is the long channel counterpart of drain-induced barrier lowering (DIBL), which is a second order effect in long channel devices fabricated in an advanced technology, that pocket implants are used at the ends of the channel for reverse $V_{TH}$ roll-off. In these devices, the pocket implants, with a higher doping concentration, will invert at a higher voltage than the actual channel region, again, assuming the operation of an NMOS. However, a high drain voltage now can modulate the barrier presented by this pocket implant, and connect the drain side to the channel before the assumed threshold voltage is reached. Subthreshold current, is very sensitive to the threshold voltage by its governing equation. One should take extra caution on the model quality for GITS, if accurate simulation results is necessary.
Figure 2.2: A measured result of deep-subthreshold current compared to simulation data.

However, both of the two mechanisms can be taken under control by limiting the maximum amount of drain-source voltage allowed in a design. This can be done simply through the usage of a lower supply voltage, or a careful stacking of devices.

Another set of modeling problem may arise depending on the technology vendor’s (or fab’s) offered capability in the extraction of parasitics. Aside of capacitor extraction accuracy issue that also annoys high frequency designers, the accurate extraction of well proximity effects and parasitic diodes are problematic in unique ways for subthreshold designs. The well proximity effect changes the effective mobility as well as the threshold voltage for the devices affected. By a careful layout, the designer can use this to further decrease the leakage current. However, the exact amount of benefit heavily depends on the accuracy of the extracted netlist. Parasitic diodes, although usually reverse biased, will introduce an additional parasitic current, usually between the supply rails.
Sometimes, this leakage current will add into the signal path, if triple-well or active bulk biasing is used. Careful design can help eliminate their contribution to the signal path in differential mode operations, but the bias current in fully differential current appears as a common mode signal, and thus needs to account for some extra margin. Multiple layer of guard rings can also be placed carefully to minimize the tiny extra power consumption it may introduce. These structures are used primarily in circuits that has direct connection to interface I/O pins for ESD related latch-up prevention [18, 19, 20].

2.1.2 Simulating MOSFETs in Deep-Subthreshold

On the simulator side, possibly the most important thing is to have a reasonable $g_{\text{min}}$ in the settings. Too large a $g_{\text{min}}$ will easily present as a short to the ground. Consider a chain of one thousand inverters under a $V_{DD} = 1$ V power domain, a $g_{\text{min}}$ of $10^{-12}$ S (default value) will give rise to 500 pA extra current consumption as a simulation artifact in terms of static power, if the $G_{\text{OFF}}$ of the inverter is considered as negligible. Furthermore, the resistive load from $g_{\text{min}}$ can be significant enough to interfere with key metrics like delay, output voltage level, static power consumption, etc. A $g_{\text{min}}$ of 0 is certainly a choice, but this may cause convergence issues from time to time in large circuits. Within the power level discussed in this work, a $g_{\text{min}}$ of $10^{-18}$ S was found as a good empirical value. In the cases where convergence issue arises in .dc simulations, a higher $g_{\text{min}}$ can be used to accelerate this process, if only .tran simulation result in steady state is of the designer’s interest.

Simulators should be used with extra care if they are capable to perform automatic MOSFET model reduction for simulation acceleration.

2.2 Deep-subthreshold Analog Design

Analog design can be massively simplified due to the elegancy introduced by Equation 2.3. The most well known one is probably the following statement: when $V_{DS}$ is higher than $4\phi_T \approx 100$ mV (assuming room temperature), the dependence of $I_D$ on $V_{DS}$ becomes negligible. This is
observation is sometimes referred to as “subthreshold saturation”, where a relatively low \( V_{DS} \) is good enough to generate an extremely low \( g_{DS} \) (or a high \( r_O \). The author of this work strongly prefer the usage of \( g_{DS} \) over \( r_O \). Similarly, \( g_m / g_{DS} \) over \( g_m r_O \). And this low value of \( 4\phi_T \) drain source voltage required allows more transistors being stacked than what strong inversion devices can afford.

The subthreshold saturation approximation mentioned above is just one of the simplifications that operating in deep-subthreshold can bring. The rest of this section will offer a rough overview on several important observations that can lead to a simplified analog design methodology for non-performance-critical deep-subthreshold analog circuits.

2.2.1 Transconductance Efficiency and Intrinsic Gain

Quantitative relationship between key transistor parameters offers another aspect of simplification in the design process. It can be better appreciated, if an inversion coefficient based design flow is adopted [21, 22]. At an extremely low inversion coefficient, the transconductance efficiency stays in its upper limit, which can be derived as follows, using Equation 2.3:

\[
\frac{g_m}{I_D} = \frac{1}{n\phi_T} \tag{2.5}
\]

A constant transconductance efficiency \( g_m / I_D \) means that the small signal transconductance is only a function of the drain current. This will lead to useful results discussed later in this section. A similar relationship can be derived for \( g_{DS} \), and hence \( g_m / g_{DS} \):

\[
\begin{align*}
g_{DS} &= \frac{dI_D}{dV_{DS}} = \frac{1}{\phi_T} I_D e^{(V_{GS}-V_{TH})/n\phi_T} e^{-V_{DS}/\phi_T} \\
&= \frac{1}{\phi_T} I_D \left[ (1 - e^{-V_{DS}/\phi_T})^{-1} - 1 \right] \\
&\approx \frac{1}{\phi_T} I_D e^{-V_{DS}/\phi_T}
\end{align*}
\]
Figure 2.3: Simulated transconductance efficiency $g_m/I_D$ and intrinsic gain $g_m/g_{DS}$ for a thick oxide 3.3 V NMOSFET.

$$\frac{g_m}{g_{DS}} \approx \frac{1}{n} e^{V_{DS}/\phi_T}$$  \hspace{1cm} (2.7)

This equation predicts an exponentially increasing intrinsic gain, that is only a function of $V_{DS}$ and not a function of sizing. Although approximation is used to derive the expression for the intrinsic gain of a MOSFET, the larger $V_{DS}$ is, the better this approximation is if the behavior of the MOSFET is fully governed by Equation 2.3 Thus with a high enough $V_{DS}$, the relationship between $g_m/g_{DS}$ and $V_{DS}$ should be closer to exponential. But this result is not observed from simulating the model of a MOSFET.

In Figure 2.3, simulation results are plotted for both the transconductance efficiency $g_m/I_D$ and the intrinsic gain $g_m/g_{DS}$. Red shade is used to indicate the region commonly used for deep-subthreshold analog design. The constant $g_m/I_D$ approximation agrees well with the simulation data, while a decrease in the transconductance efficiency at lower drain current level (Figure 2.3 (A)) is considered as an artifact from modeling and simulation. On the intrinsic gain side, not only does simulation shows an increasing intrinsic gain with longer channel length, but the intrinsic
gains’s growth slows down immediately after entering subthreshold saturation ($V_{DS} > 100$ mV), as shown in Figure 2.3 (B). At this point, it is unclear how much of this discrepancy belongs to modeling issues, and how much from the effects modeled, but not captured in Equation 2.1.

2.2.2 Speed Consideration

This subsection studies the speed of closed-loop amplifiers that are built with MOSFETs biased in deep-subthreshold. The feedback network is usually a switched capacitor one. This is because a MOSFET in deep-subthreshold saturation has a low $g_m$ (assuming a source follower output stage), and a degrees of magnitude lower $g_{DS}$ (assuming other kind of output stage). Any resistors implemented using a reasonable area will load the output significantly. Now assume the feedback ratio of the network is $\beta$. For a stable amplifier, for simplicity, a single dominant pole is assumed in its frequency response. Then the unit-gain bandwidth can be written as:

$$\omega_U = \frac{g_m}{C_P},$$

where $g_m$ is the first stage transconductance, and $C_P$ is the capacitor that defines the dominant pole. It is the load capacitance for a single stage amplifier, or the miller capacitance for a two stage amplifier. Since the closed loop gain is set by $\beta$, and its 3dB bandwidth is set by:

$$\omega_{3dB,CL} = \beta \omega_U = \frac{\beta g_m}{C_P},$$

and with $g_m$ being set by a constant $g_m/I_D$ times the drain current $I_D$. The above observations yields a simple, yet “rigid” design guideline. With a specified signal gain (this is often equal to the closed loop gain), and the bandwidth of the signal, a linear increase in the capacitor added to the load gives a linear increase to the $I_D$, that yields a linear increase in power. Note that in the case of a miller compensated scheme, a linear increase in the output load capacitor $C_L$ will move the location of the second pole, thus a increase in the miller capacitor $C_C$ is then required to keep the same phase margin for stability considerations.
However, current alone does not determine the power. But the choice of supply voltage comes from a different perspective.

2.2.3 Noise Consideration

Under deep-subthreshold operation, noise is very likely dominated by the flicker noise of the transistors, since the frequency of operation is usually below 10 kHz. Thermal noise can, however, play an important role in switched-capacitor topologies, as sampling brings noise folding right during the operation. Here, we assume the amplifier has its own poles in the transfer function, where the total integrated noise can be brought down into a scaled $kT/C$ form. In fact, the exact closed form of noise power in an amplifier is still not so well studied, as simple circuits can easily create complicated analytical solutions, attracting even the most recent efforts for an intuitive understanding [23, 24]. In this thesis, we avoid the usage of an exact solutions, but assumes a practical situation where simulated noise has exceeded the design requirement, and investigate how the noise can be reduced from sacrificing other metrics, i.e. trade-offs between other design parameters.

The input-referred flicker noise voltage in a transistor is often theoretically predicted and modeled analytically as [16]:

$$S_{vf}(f) = \frac{K_F(V_{GS})}{C_{OX}^2 W L f^c}. \quad (2.10)$$

Within deep-subthreshold, we consider the factor $K_F$ a constant, as variation in $V_{GS}$ is usually small. Thus the only way to reduce device’s flicker noise, is to increase its width (length fixed due to intrinsic gain consideration). If the transistor of consideration is being driven by a similar analog circuit implemented in deep-subthreshold, and its gate capacitance presents as the dominant load, the current consumption in the previous stage needs to be increased accordingly, for a higher $g_m$, to maintain the same bandwidth. However, if an additional passive capacitor is the dominant load for the previous stage, then sizing such transistor up only has area penalty, until it presents as a significant load. Thus with other design parameters fixed, a linear reduction in the flicker noise lead to a linear increase in area, and at most, a linear increase in power consumption.
Another method, possibly more well known when discussing analog scaling, is to increase the signal swing, which usually requires an increase in the supply voltage $V_{DD}$ [25, 26]. The idea is to keep the noise the same, while increasing signal level, to effectively boost the signal-to-noise ratio (SNR). This, however, require a strict linear increase in power consumption, but comes at no area penalty.

2.2.4 Mismatch Consideration

Another metric to consider is the mismatch. Effects and controls for mismatch in threshold voltage $V_{TH}$, device width $W$ and device length $L$ will be discussed here. On top of that, we focus on two scenarios: mismatch effect in a current mirror, and that in an input differential pair.

The mismatch effect is usually modeled as the variance for $V_{TH}$, and relative variance for $W$ and $L$, in the following way related to the geometric factor $1/\sqrt{WL}$:

\[
\sigma_{VTH} = \delta V_{TH} \sqrt{WL} \quad \text{(unit: mV} \cdot \mu\text{m},) \\
\sigma_W = \frac{\delta W}{W} \sqrt{WL} \quad \text{(unit: } \mu\text{m}),
\]

and
\[
\sigma_L = \frac{\delta L}{L} \sqrt{WL} \quad \text{(unit: } \mu\text{m}).
\]

(2.11)

Writing the exact value down would be a violation to the non-disclosure agreements, but to get a sense of how much each of these contribute to the final mismatch, we use $\sigma_{VTH} = 3$ mV·µm, and $\sigma_W = \sigma_L = 0.2\%\cdot\mu$m. These value are simply used for quantitative comparison of the contribution of each mismatch mechanism, and the conclusion arrived here subjects to changes from the actual technology’s mismatch parameters.

For a current mirror, the same $V_{GS}$ is shared among a set of devices. It is easy to see that $W$ and $L$ contributes directly to the variance of the relative drain current:

\[
\frac{\delta I_{D,W}}{I_D} = \frac{\delta W}{W} = \sigma_W \sqrt{WL},
\]

\[
\frac{\delta I_{D,L}}{I_D} = -\frac{\delta L}{L} = -\sigma_L \sqrt{WL},
\]

(2.12)
However, the contribution from $V_{TH}$ is:

$$\frac{\delta I_{D, VTH}}{I_D} = -\frac{1}{n\phi_T} \delta V_{TH} = -\frac{g_m}{I_D} \frac{\sigma_{VTH}}{\sqrt{WL}}.$$  \hspace{1cm} (2.13)

Equation 2.5 is used to derive the result in Equation 2.13. In deep-subthreshold, $g_m/I_D$ can be very well considered as a constant. Here we use $g_m/I_D = 30 V^{-1}$ as an estimation for CMOS. With all estimated value plugged into the expressions, it is easy to see that the relative variation of $I_D$ is more sensitive to that in the variation of $V_{TH}$, at a level of 9%·µm. The above calculation, however, assumes the transistor that generates this $V_{GS}$ does not suffer from mismatch of any kind. This is not true, but to take this into consideration, a simple factor of $\sqrt{2}$ needs to be multiplied to the above derived value, as the best assumption to apply here is the mismatch of different devices follows independent gaussian distributions. This elevates the previous calculated number to 0.28%·µm for $W$ and $L$ mismatch, and 12.7%·µm for that from $V_{TH}$ mismatch. Thus, to control for accuracy in current copying operations, the designer can simply use the ratio between the desired accuracy and the relative contribution of $V_{TH}$ mismatch from a unit area transistor, to find the minimum required area.

If this mismatch is happening in an input differential pair, the situation is slightly different, as what is most interesting, now, is to find the input offset of this differential pair. This can be found by assuming the two transistors passing the same drain current, biased in deep-subthreshold saturation.
(ignoring the $V_{DS}$ term), but having different $V_{GSS}$:

\[
I_{D1} = I_{D2}
\]

\[
I_S \frac{W + \delta W_1}{L} e^{(V_{GSS1-V_{TH1}})/n\phi_T} = I_S \frac{W + \delta W_2}{L} e^{(V_{GSS2-V_{TH1}})/n\phi_T}
\]

\[
V_{GS1} - V_{GS2} = n\phi_T \ln \left(1 + \frac{\delta W_2}{W}\right) - n\phi_T \ln \left(1 + \frac{\delta W_1}{W}\right)
\]

\[
\approx n\phi_T \left(\frac{\delta W_2}{W} - \frac{\delta W_1}{W}\right)
\]

\[
= \frac{1}{g_m/I_D} \sqrt{\frac{1}{WL}} \left(\sigma_{W2} - \sigma_{W1}\right)
\]

\[
\sigma_{VOS} \approx \frac{1}{g_m/I_D} \sqrt{\frac{2}{WL}} \sigma_W
\]

(2.14)

The last line from the Equation 2.14 considers that the random variable $\delta W_1$ and $\delta W_2$ are uncorrelated, and follow the same gaussian distribution. Plugging in the numbers from our assumed technology, this gives a 0.094 mV·µm offset voltage. It can be derived in a similar if not identical process, that the mismatch in $L$ contributes to the offset voltage in the same fashion. However, that from the threshold voltage is different:

\[
I_{D1} = I_{D2}
\]

\[
I_S \frac{W}{L} e^{(V_{GSS1-V_{TH1}})/n\phi_T} = I_S \frac{W}{L} e^{(V_{GSS2-V_{TH2}})/n\phi_T}
\]

\[
V_{GS1} - V_{GS2} = V_{TH2} - V_{TH1}
\]

\[
= \frac{1}{\sqrt{WL}} \left(\sigma_{VTH2} - \sigma_{VTH1}\right)
\]

\[
\sigma_{VOS} \approx \frac{\sqrt{2}}{\sqrt{WL}} \sigma_{VTH}
\]

(2.15)

And this gives a one sigma of 4.24 mV·µm in the offset voltage. Again, we can see that the offset voltage is dominated by the mismatch from the threshold voltage. Thus in a design process, a designer can, again, directly estimate the area needed for the input transistor from the desired offset voltage.

It is worth noting that the effect of oxide thickness $t_{OX}$ mismatch is also commonly modeled, but
not discussed here. This is because the oxide thickness affects many physical parameters indirectly, and its contribution is hard to detail. The accuracy of certain parameters is also in question for deep-subthreshold designs. The relative mismatch found in oxide thickness is usually at the same level as that in device dimensions, and its direct contribution to $C_{OX}$ is similar to the variation in $L$, which has been proven to be negligible.

Another interesting fact is that within the calculation that has been performed here, the sensitivity of the desired performance (current matching or input referred offset) to threshold voltage mismatch is always a $g_m/I_D = 30 \, V^{-1}$ higher than that to the geometry. The reader may also recall that in the textbook [6], a similar derivation is done in strong inversion. And the quantitative relationship between different mismatch mechanisms look identical to what we just described, if $2/(V_{GS} - V_{TH})$ is replaced with $g_m/I_D$. In the derivation presented here, as well as that in [6], the conclusion uses the explicit expression of the drain current. However, in fact, this conclusion can be generalized if the transistor’s drain current expression meets the following requirements:

1. Geometric term $W$ and $L$ only show up in the drain current as a proportional factor $W/L$,

2. Gate voltage $V_{GS}$ and threshold voltage $V_{TH}$ only appears in the form of $(V_{GS} - V_{TH})$.

We can derive, from assumption No. 1:

$$I_D = \frac{W}{L} I_0$$

$$\frac{dI_D}{d(W/L)} = I_0 = \frac{I_D}{W/L}$$

(2.16)

Then:

$$\delta I_D = \frac{dI_D}{d(W/L)} L \delta W = \frac{I_D}{W/L} L \delta W$$

$$\frac{\delta I_D}{I_D} = \frac{\delta W}{W} = \frac{\sigma_W}{\sqrt{WL}}$$

(2.17)
\[
\delta I_D = \frac{dI_D}{d(W/L)} \left( -\frac{W}{L^2} \right) \delta L = -\frac{I_D}{W/L} \left( \frac{W}{L} \right) \frac{1}{L} \delta L
\]
\[
\frac{\delta I_D}{I_D} = - \frac{\delta L}{L} = -\frac{\sigma_L}{\sqrt{WL}}.
\]

(2.18)

And from assumption No. 2:

\[
g_m = \frac{dI_D}{dV_{GS}} = \frac{dI_D}{d(V_{GS} - V_{TH})} \frac{d(V_{GS} - V_{TH})}{dV_{GS}} = \frac{\partial I_D}{\partial (V_{GS} - V_{TH})}
\]
\[
\delta I_D = \frac{\partial I_D}{\partial (V_{GS} - V_{TH})} \frac{d(V_{GS} - V_{TH})}{dV_{TH}} \delta V_{TH} = -g_m \delta V_{TH}
\]
\[
\frac{\delta I_D}{I_D} = - \frac{g_m}{I_D} \delta V_{TH} = -\frac{g_m \sigma_{V_{TH}}}{I_D \sqrt{WL}}.
\]

(2.19)

This gives a generic derivation of the sensitivity to different mismatch mechanism in MOSFET-based analog circuits operating in all regions. Depending on the actual variation in each parameter, and the inversion coefficient, which determines the transconductance efficiency \(g_m/I_D\), one or more mismatch mechanisms may be neglected in the design consideration. If different mismatch mechanisms are uncorrelated, then they should be added in a root-mean-squared way. If one source of mismatch contributes to less than 20% of the dominant source of mismatch, then it only contributes to about 2% in the standard deviation of the final total mismatch \((\sqrt{1 + (1/5)^2} = 1.0198)\).

The most important aspect of this discussion, is that by using the above conclusion, mismatch can be brought into consideration in an early design stage. And if a dominant mismatch source exists in the circuit topology, the metric of interest can be accurately estimated from reading the model files and finding each mismatch mechanism’s standard deviation, without invoking complicated Monte Carlo simulation.

2.2.5 Operational Transconductance Amplifier Design Flow

A simple operational transconductance amplifier (OTA) design flow is presented in this section to better demonstrate how discussions in the previous subsections should be applied in systematic deep-subthreshold operational amplifier design.
Assuming an amplifier is used in a switched capacitor feedback loop, and its output swing, unit-gain bandwidth, dc gain, input voltage offset, and total integrated noise has been identified from closed loop gain, closed loop output signal swing, closed loop 3dB bandwidth, and desired output SNR. To choose a proper device length, a sweep for $g_m/g_{DS}$ against different $L$ needs to be performed, and several fixed length-intrinsic gain pairs can be determined. Output swing and dc gain together can then help the designer to choose a suitable topology. Start with the effective load capacitance, the minimum $g_m$ can be derived from Equation 2.8, and bias current at the $g_m$ stage can be set accordingly, from Equation 2.5. Read the model file for the $1\sigma$ mismatch in $V_{TH}$, and upsize matched transistors in the multiplication factor until a desired mean mismatch is achieved. If flicker noise becomes a concern, continue to upsize until it is not anymore. If thermal noise start to matter in rare cases, like when significant frequency folding happens, one can simply upsize both the entire circuit with current source, as well as load capacitor, until a desired noise floor is achieved (a process introduced at least in [6]).

What this oversimplified design flow brings, is a fast navigation within a very confined design space, surrounded by “rigid walls”: little to no resistance is met and designer’s efficiency is maximized when designing within the allowed performance metrics, but walking out of it is extremely hard. Metrics like linearity, slew rate, PSRR, and CMRR does not show up in the design flow, as they require a much deeper look into the actual topology and/or application scenario. Topologies either more complicated or more simplified that achieve better performance exists, for instance [27], but poses themselves new sets of trade-offs.

2.2.6 Summary and Discussion

This section provides an oversimplified walkthrough of analog design in the deep-subthreshold regime. The findings here, are by no means novel and have been observed by generations of low power analog designers from different perspectives, for instance [25] and [28]. Here, methods are rearranged in the modern inversion coefficient “syntax”, with a highlight on the design perspective. On top of the above disclaimer, there are at least two other important topics that are left out in the
discussion presented in this section.

The first one is the usage of switch based techniques for mismatch and flicker noise reduction. This includes auto-zero, correlated double sampling, and most importantly, chopper stabilization. These techniques are probably best summarized in [29]. Adopting such techniques, especially chopper, brings a massive reduction of flicker noise at the most noticeable price of a relatively reasonable frequency clock. Using them, a designer can take advantage of a much improved noise performance to reduced power consumption or area utilization of the circuit while keeping other metrics mostly unchanged. Even nowadays, these techniques still enjoy frequent show-off in low power amplifier designs achieving record performance [30, 31, 32]. However, the work involved here was not able to use chopper due to its specific application scenario, which will be introduced in detail in Chapter 4. Thus introduction to these techniques is omitted in this thesis, though they certainly serves an important role in modern low power analog design.

The second one is the usage of bipolar junction transistors (BJT). Even nowadays in industry, a sizable portion of analog IC’s are still designed using hybrid BJT and CMOS, or fully in BJT. BJT is most well known for achieving the highest possible transconductance $g_m/I_D = 1/\phi_T \approx 38.6 \ V^{-1}$ during operation, which, from the discussion earlier, is certainly more advantageous than its MOS friend. However, the biggest reason to avoid BJT’s in this work, is for its base current. This base current will add an extra trade-off on the size of the feedback capacitor in a switched-capacitor amplifier, as now, a tiny leakage current is present in the presumably charge conserved node: the differential input of the OTA. However, usage of BJT is certainly justified elsewhere to achieve a hybrid design, a possibility the author of this work unfortunately didn’t well explore.

2.3 Low-Voltage Digital Design

Power consumption of a digital circuit can be brought down into two parts. The dynamic power, the part of energy dissipated for information processing, and the static power, the part that will be consumed as long as the digital circuit is powered up. Since the information is stored as charge on a capacitor in CMOS digital circuits, dynamic power is largely related to how much capacitance
needs to be charged, and how high an voltage the capacitor needs to be charged to. Nothing reduces the total amount of capacitance better than a smaller feature size and a simpler topology, limiting the possible variations in the technique for dynamic power reduction. Complementary CMOS with occasional assist from transmission gate logic thus provides elegant and simple implementation of logic elements, and is often the go-to solution for the implementation of digital systems. However, with low frequency operations, it is also possible to trade dynamic power for a lower leakage, as the total power consumption can be oversimplified into:

\[ P_{TOT} = E_{DYNS}f_{CLK} + P_{STA}. \]  

(2.20)

Thus for a certain operating frequency, the optimal balance for dynamic power and static power may shift. This inspires a set of new logic gate topology for low frequency digital circuits. This section will include one notable topology that uses the super-cut-off operation for leakage reduction, presented initially under the name “dynamic leakage suppression” (DLS) logic [33].

Finally, all discussions here focused on a fixed, “low”, clock frequency in the sub-100 Hz range, which applies to all design scenarios within the scope of this thesis (Chapter 3 and Chapter 4). Leakage suppression is more important than dynamic power reduction for such a clock frequency when using 180 nm design package. The optimization process presented here is sufficient for the scope of this work, but it cannot be generalized to all design scenarios, since minimizing energy per operation will lead to a different but overlapping set of considerations. Some conclusions translates, yet certain others will not hold anymore, breaking the reasoning that is the backbone for the optimization process introduce in this chapter.

2.3.1 Complementary CMOS in Deep-Subthreshold

Many considerations regarding deep-subthreshold CMOS logic design has been introduced by a variety of publications. Most design considerations presented here are similar to what is described in [34], but was also partly found from a different perspective presented in [25].
Consider a simple 2 stage CMOS inverter chain shown in Figure 2.4, where currents are labeled for the two cases when input is a logic “0” and a logic “1”. The output voltage, $V_{OUTS}$ in the two stages, are determined by:

$$I_{OFF,N}(V_{DS} = V_{OUT,1}) = I_{ON,P}(V_{DS} = V_{DD} - V_{OUT,1})$$

$$I_{ON,N}(V_{DS} = V_{OUT,2}) = I_{OFF,P}(V_{DS} = V_{DD} - V_{OUT,2})$$

(2.21)

It is indeed possible to solve such equations to find an analytical solution to this problem, and then determines the choice of supply voltage $V_{DD}$ and sizing. One example of such calculation is done in [34], where the behavior around $V_{DS} = V_{DD}/2 = V_M$ is investigated. From their analysis, inverters will lose its gain at $V_M$, when $V_{DD}$ is set to below $2\phi_T$. They also extend their discussion to process variation induced threshold voltage shift, and how this translated to a further elevated minimum $V_{DD}$ depending on the process variation model. Here, we instead, propose an empirical work flow in light of this theoretical discussion. In this work, little analytical results are derived to justify the work flow that will be presented later, yet the proposed empirical method yields a reasonable result without too much design efforts. It also takes several other factors into account that are usually hindered by the approximations adopted in [34], the majority of which come from the fact that the input voltage cannot be approximated as an ideal “0” (i.e., ground level) or an ideal “1” (i.e., supply voltage) when the gate is driven by logic components of the same type.

The first, relative simple observation is the impact of device sizing. Before delay becomes a
concern, minimizing width $W$ is a dominate choice, as it reduces both static leakage and dynamic power. Device length, however, offers a trade-off between an almost linear reduction in leakage current with an almost linear increase in dynamic energy. The gate area $W \times L$ also affects the choice of minimum $V_{DD}$ through mismatch, which will be discussed shortly, and most of this adjustment goes to device length for a lower leakage power. When an initial choice of $W$ and $L$ are taken, its sizing ratio can be approximately determined as the inverse of the ratio of their drain current at $V_{DD}/2$ (an initial guess of $V_{DD}$ might be needed) for delay balancing.

The second consideration from this observation is the fact that “stuck at 0” and “stuck at 1” inverters may happen under fast-NMOS-slow-PMOS (FS) process corner or slow-NMOS-fast-PMOS (SF) corner. When the threshold voltage of an NMOS and a PMOS shifts to opposite directions, $V_M$ may drift away significantly from its desired value of $V_{DD}/2$, to a level not able to invert the next stage. It is easiest to observe this effect in a long inverter chain, as the output voltage of every other stage slowly converged to two certain values, i.e., $V_{OH}$ and $V_{OL}$. A minimum $V_{DD,MIN,PROCESS}$ can be derived from a sweep in the supply voltage across all process corners, to find when neither the “stuck at X” problem will occur. However, from a similar principle, an adjustment to this must be made for local mismatch (see Subsection 2.2.4). With a large gate count in a typical digital system, $3\sigma$ or even $6\sigma$ variation should be compensated for a negligible impact on the yield, leading to a new, adjusted $V_{DD,MIN}$. However, since mismatch scales with device area, this step must be done iteratively when resizing happens. It is important to note that a change in device dimension also changes the threshold voltage due to its dependency on $W$ and $L$, leading to a different $V_{DD,MIN,PROCESS}$. Furthermore, in certain cases, around $V_{DD,MIN}$, the leakage current is non-negligible under certain process corner. This is because even the output of the inverter does not stuck at “0” or “1”, it may still have a non-negligible difference from the rail, causing an exponential increase in the subthreshold leakage. A further increase in $V_{DD}$ may be beneficial for overall power reduction.

To summarize, the above qualitative analysis can be put into an empirical, iterative work flow for leakage power optimization using a long inverter chain. An even stage inverter ring can also
be used with a proper initial condition (nodeset for dc simulations). Starting with minimum sized transistors, the $V_{DD,MIN,PROCESS}$ can be first derived by simulations across different corners (simulating both FS and SF is usually sufficient). The minimum $V_{DD,MIN}$ is then further adjusted together with an increase in device length $L$, for a reduced impact from mismatch to the overall yield by running Monte Carlo simulations. A proper $V_{DD}$ can then be chosen iteratively with adjusted device sizing for the minimum leakage power.

2.3.2 Dynamic Leakage Suppression Logic

Since its first introduction 5 years ago [33], DLS logic has seen its expanding usage in applications from traditional computation unit [35] to selective applications in low power, low frequency IoT devices [36, 37]. An example DLS inverter is presented in Figure 2.5, (A), as well as its simulated time domain behavior under fan-out-of-four (Fo4) configuration. The principle of operation of an DLS inverter can be explained as following.

![Figure 2.5: Schematic of a DLS inverter, and its simulated time domain waveform under Fo4 configuration.](image)

The PMOS $M_{PI}$ and the NMOS $M_{NI}$ in the middle makes a complementary CMOS inverter,
however, their access to power rails are further gated by another set of transistors $M_{NT}$ and $M_{PT}$.

When $V_{IN}$ is low, node $V_{MP}$ can be seen as approximately shorted to $V_{OUT}$. $M_{NT}$ is then charging node $V_{OUT}$ at a maximum current of its subthreshold leakage current $I_D(V_{GS} = 0V)$. $M_{NI}$ is now off, and will push node $V_{MN}$ down until it passes the same subthreshold leakage as $M_{PT}$. With proper sizing making the leakage through $M_{NT}$ ($V_{GS,NT} \approx 0V$) larger than that through $M_{NI}$ ($V_{GS,NI} \leq 0V$), $V_{OUT}$ is guaranteed to be pulled up, further cutting off transistor $M_{PT}$. Thus the equivalent resistance of the upper two transistors ($M_{PI}$ and $M_{NT}$), limited by the subthreshold leakage of $M_{NT}$ under zero $V_{GS}$, is lower than that of the bottom two transistors ($M_{NI}$ and $M_{PT}$), where both of them are under super-cut-off ($V_{GS} < 0V$). The output voltage is pulled high. Similarly, when input is high, the output is low, and this block implements an inverter as what is shown in Figure 2.5 (B) from simulation.

From the discussion above, it is also straightforward to see that the speed of a DLS inverter is primarily limited by the subthreshold leakage of the “gate” transistor $M_{NT}$ and $M_{PT}$ in the “on” side, while an exponential reduction of leakage current can be achieved by a negative $V_{GS}$ seen on the “off” side. Since the magnitude of the negative $V_{GS}$ is proportional to the supply voltage, a higher $V_{DD}$ will reduce the leakage current before GIDL becomes observable. Sizing of the “gate” transistors can be optimized by balancing speed and leakage for a desired operating frequency. The delay of a DLS circuit is usually in the range of milliseconds, making them useful mostly in sub-100 Hz clock frequencies.

One major speed limitation in DLS circuits comes from its intrinsic feedback mechanism: the signal at the output controls the super-cut-off tail devices, which in turn, generates the very output signal. This feedback limits the $V_{GS}$ of the gate transistors (for simplicity without loss of generality, the discussion here focuses on the NMOS gate transistor $M_{NT}$) from reaching above 0 V. However, if the input logic comes with its inverse, then a feedforward topology can take advantage of this to generate a positive $V_{GS}$, accelerating the discharging of the output node. This possibility is investigated by [38, 39], where a fully differential logic (feed-forward leakage suppression logic, or FLSL) is implemented achieving extremely low leakage power with > 1 kHz operating frequency.
However, such topology intrinsically requires a higher amount of capacitance being charged for the same logical function, as it needs 2 times more transistor counts per gate, and wider “gate” transistors to ensure a proper functionality and a reasonable performance. Its dynamic power consumption hardly justifies the operating frequency.

Although the usage of the DLS and its derivatives meets a small application scenario space, its design, and the usage of the super-cut-off pair ($M_{NT}$ and $M_{PT}$, as well as $M_{NI}$ and $M_{PI}$) give unique inspiration for leakage suppression, and can be used with modifications when the leakage current of the transistor plays a key role.

### 2.4 Building Blocks for Sub-nW Systems

This section details the design of two building blocks that are shared across different deep-subthreshold designs. This includes a close-to-minimum standard cell library for the implementation of almost all the digital logic that will be described later, and an ESD protected I/O ring including a customized low leakage power clamp for sufficient ESD protection with a negligible static power consumption.

#### 2.4.1 Standard Cell Library

One obstacle in implementing a fully integrated sub-nW system, is that most of the standard cell libraries available in the market are optimized targeting for a high performance but not a low power. Even low-power favored libraries are likely to induce pico-amps of leakage per gate from commercial standard cell vendors. Simple logic blocks can easily use thousands of logic gates when synthesized, and this level of leakage is rarely acceptable in sub-nW designs. Thus a custom built standard cell library is necessary, for at least a lower static leakage current per gate.

To do this, a custom optimized standard cell library is implemented in traditional complementary CMOS logic gate topology using I/O (3.3 V) standard threshold devices. It achieves an overall smaller area than the DLS logic implemented using 1.8 V core devices, and a negligibly elevated leakage thanks to an intrinsically higher threshold voltage. With a low supply voltage, below or
near threshold, the “on” current for the logic gates are generally low to a point where I-R drop on the power rail becomes negligible. Thus in layout, power delivery related considerations can be ruled out during power grid generation, and space can be saved by using narrower, even close-to-minimum power traces. The example of an inverter (1XINV) and that of a two input NAND gate (1XNAND2) are shown in Figure 2.6. The complete list of cells in this standard cell library can be found in Table 2.1. The leakage current is dominated by the subthreshold leakage of the I/O devices, which is 1.8 fA in average across different input values for the inverter 1XINV under typical-typical (TT) process corner.

Figure 2.6: The layout of an inverter and a NAND gate in the standard cell library.

This standard cell library (named O3V for optimal-3.3 V) is characterized across a supply voltage $V_{DD}$ between 0.5 V to 1.0 V, all $3\sigma$ process corners, and the temperature range of interest for implantable applications. Synthesized logic can operate up to 1 kHz a clock frequency safely. It is uses to synthesis all digital logics in the future chapters (Chapter 3 and Chapter 4), with possible
minor modifications in the layout between different projects.

<table>
<thead>
<tr>
<th>INV-(1X-4X)</th>
<th>Inverter</th>
<th>BUF-1X</th>
<th>Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND2-1X</td>
<td>2 Input NAND</td>
<td>NOR2-1X</td>
<td>2 Input NOR</td>
</tr>
<tr>
<td>XOR2-1X</td>
<td>2 Input XOR</td>
<td>XNOR2-1X</td>
<td>2 Input XNOR</td>
</tr>
<tr>
<td>MUX2-1X</td>
<td>2 Input MUX</td>
<td>BUFT-1X</td>
<td>Tri-State Buffer</td>
</tr>
<tr>
<td>DFFR-1X</td>
<td>D-Flip-Flop with Async-Reset</td>
<td>DFFS-1X</td>
<td>D-Flip-Flop with Async-Set</td>
</tr>
<tr>
<td>LATCHR-1X</td>
<td>Latch with Async-Reset</td>
<td>TIE-(HI</td>
<td>LO)</td>
</tr>
</tbody>
</table>

Table 2.1: Available cells in the standard cell library (O3V).

2.4.2 Low Leakage ESD

Another obstacle in implementing a complete sub-nW system comes from the ESD structure. ESD’s are used to protect the core circuit from events that generate large voltage spikes across I/O pins of an IC. These events are commonly found when the IC is accidentally connected to external capacitors that has a built-up static charge, like human bodies, machines, and/or charged devices. These voltage spikes usually have a limited amount of total charge, but the voltage induced on the IC can reach kilo-volts in less than a nanosecond, until a discharge path is found. The high voltage may cause physical damage to the internal circuit, as an undesired discharge path can be created through breaking down the oxide, or dopant redistribution. ESD structures are designed to redirect this current away from the core circuit, when an ESD event was detected.

To avoid ESD structures interfering with normal circuit performance, one of the most widely used structures for ESD event protection is designed as follows. Each I/O pad is protected by a primary ESD device, that provides a discharge path to the high power rail $V_{DD}$, and a discharge path from the low power rail $V_{SS}$. A specially designed power clamp is placed to connect $V_{DD}$ to $V_{SS}$, which only turns on if it detects a high voltage jump with fast transient appears on the $V_{DD}$, a typical characteristic of an ESD event. Figure 2.8 provides a simplified overview of the structure described above. A more detailed description with other alternative structures can be found in at least [19].

32
Commercial ESD structures, although silicon-proof, introduces a non-negligible static leakage between $V_{DD}$ and $V_{SS}$, mainly due to the leakage from the power clamp circuit. A measured leakage current of a commercially available ESD structure is shown later in Figure 2.10 for comparison purposes. A 10 nA leakage is not rare for such devices, as this provides negligible power overhead to most of the commercial ICs. However, such a leakage is overwhelming if the system is designed to operate under sub-nW range. It is also worth noting that although low power ESD structures exist, they may rely on novel device structures that circuit designers don’t have easy access to, like the one presented in [40].

![Schematic of the proposed ultra-low-leakage power clamp.](image)

Figure 2.7: Schematic of the proposed ultra-low-leakage power clamp.

To reduce the leakage in the power clamp during normal operation, a novel ESD power clamp structure is proposed, which is shown in Figure 2.7. The key is to replace the conventional discharging transistor, which is a single NMOS, with a super-cut-off pair of NMOS and PMOS stack. When the two transistors are placed in series, their width needs to be at least doubled to achieve the same discharging capability. Other blocks, including the supply line high pass filter and inverters, stays the same as what is commonly used. A capacitor $C_{KEEP}$ is added to reduce the transient voltage rising through capacitive coupling during the onset of the ESD event at the gate of the PMOS transistor. This placement and the sizing of this capacitor is chosen by simulation empirically, where the addition of this capacitor has shown to largely accelerate the discharging
Figure 2.8: The simulation testbench setup for the proposed ESD structure under HBM ESD events. The anticipated discharge path is highlighted with a red arrow.

The power clamp, as well as pad-level ESD protections, are assembled into a pad ring, and are placed under JEDEC standard ESD testbenches [41, 42]. The testbench for HBM is shown in Figure 2.8, and the simulated voltage at node $V_{DD}$ subject to 1 kV HBM is plotted in Figure 2.9.

It is found that the power clamp can discharge incoming ESD pulses down to safe operating voltage range within 1 $\mu$s for up to 1 kV HBM events. Simulation with an even higher ESD stress does not make much more sense. This is because when an ESD event happens, the place that is the weakest to the incoming ESD stress will become the breaking point of the entire circuit, and layout guideline provided by the foundry only protects up to 1 kV HBM for ESD induced latch-up [18]. This simulation result is not yet verified with real world experiments, as JEDEC compatible measurement setup was not available in the lab at the time. Thus the actual ESD discharging capability of this custom designed ESD structure remains a question.

Finally, the leakage current of this design, both simulated and in real measurement are plotted in Figure 2.10. The leakage current measurement result of a commercial power clamp from Taiwan Semiconductor Manufacturing Company (TSMC) is also in the figure for comparison. A greater than 450 times leakage power reduction is found comparing the proposed power clamp with the
commercial one, and a less than 10 pA leakage is achieved for supply voltage below 0.8 V. However, a significant difference between simulation and measurement is found when $V_{DD} > 0.8$ V. Although the reason is still unclear, GIDL, from the author’s perspective, remain the most likely suspect.

Efforts on sub-nW ESD structures are rarely reported. Compared to the only reference known to the author [43], this ESD power clamp achieves a higher leakage at 1.8 V nominal $V_{DD}$, yet stays much lower in leakage current when supply is less than 0.8 V.

2.5 Summary

In this section, a simplified overview of the deep-subthreshold design world is provided, covering topics from device performance to design methodology. The section includes a discussion on model accuracy, precautions in simulation, and the usage of deep-subthreshold MOSFETs in digital design as well as analog design. Two building blocks that are necessary to build sub-nW mixed signal ICs are introduced: a digital standard cell library and a pad ring embedded with low leakage ESD structures. The introduction of the two blocks above also marks the starting point of the author’s original contribution to the field.

However, the content in this chapter also suffers from great limitations. First, to avoid complications that arise from gate leakage, all the IC’s that are built in this thesis use only one technology.
Figure 2.10: Comparison of the leakage current between the proposed power clamp in simulation as well as in measurement, and that to a commercially available ESD structure.

node, the TSMC 180 nm MSRFG. Thus, the majority of the content here are based on the author’s personal experience with this specific technology node. Although the derivation here avoids the dependency on the actual technological constants, and appears to be generally well held, other physical effects (likely those that are short-channel related) that are more pronounced in deep sub-100 nm technologies may change the form of the first order trade-off curve, thus affect the first order design flow. It is important to realize that it remains an open question on how well conclusions presented here translate to other technology nodes. Second, the material presented in this chapter is also limited to the use of MOSFETs, as mentioned at the end of Section 2.2. Though recent development on solid state circuits focuses on MOSFETs almost exclusively, the author believe BJT devices should not be overlooked especially in deep-subthreshold region. After all, circuit design should fulfill system level requirements, and designers should not avoid hybrid designs if it fits the goal better. Last but not least, the design considerations presented here, though sound, only explore
an extremely limited subset of available digital and analog topologies. With a different topology, different trade-off curves can be established, and performance much better than that the simple topologies introduced here can be achieved. The derivation here should only serves as a guide, when the desired metrics of the system falls into the design space described here. The constraints presented in this chapter are not limitations to the actual possibilities of deep-subthreshold circuits.

Deep-subthreshold circuit designs can easily utilize the limited bandwidth in the source signal, and massively reduce the power consumption in the sensing process, without strict limitations (though trade-offs exists) on other important performance metrics. This, if used correctly, can be a powerful tool for bio-signal sensing applications.
Chapter 3: An Ultrasonography Compatible Implant

In this chapter, an ultrasonography compatible powering and data telemetry system is introduced, serving as a platform for next generation of miniaturized, battery-less, distributed, real-time trackable physiological signal sensing platform. Instead of using continuous ultrasound wave in a fashion similar to radio-frequency (RF) for modern communication devices, the design of the proposed sensor platform tries to be as compatible as possible to an existing and widely adopted tomography solution. This novel solution has many significant advantages that no previous works can achieve. One of them is biogeographical context aware tracking. This means the implant can be identified in real time, with information about its relative position to nearby organs, thanks to its non-interfering operating principle with the imaging session. Another significant advantage is that by its working principle, multiple implants in the same field of view can operate in parallel with little interference. Furthermore, the adoption of the proposed implant system requires minimum upgrade to existing, widely used medical equipment, as well as that in the training for front-line medical technicians. However, on the engineering side, this has imposed a set of challenges. As this chapter will discuss later, to operate the devices in a way compatible with ultrasound imaging imposes strict power limit and requires novel data communication modality.

To provide a thorough view of the proposed ultrasound sonography system, the rest of the chapter is organized as follows. The background of the problem will be introduced first, highlighting the paradox present in the previously published deep tissue implants, and the necessity to have a context aware trackable implant system. After that, the proposed solution is investigated in detail, translating the vague requirement of “sonography compatibility” into a set of practical engineering constraints, that defines the design of the frontend circuit on the system level. Circuit level implementation of each required function is then described in detail. The testing result is shown next, demonstrating the performance achieved from this prototype when operating with a minimally
modified sonography session. Finally, this chapter concludes with a brief comparison with existing work is presented and a short discussion.

This work is first demonstrated in [44].

3.1 Background

The idea of a “smart” implantable devices has attracted researchers’ attention for decades. Various of such active implants have been developed, like cochlear implants for hearing aids [45, 46, 47], implantable systems for epilepsy detection and treatment [48, 49], and brain-machine interfaces including neural activity recorders [50, 51, 52, 53, 54] and spatial specific stimulation electrodes [53, 55, 56]. The vision that an active implant is able to record, decide, and control biological signals promises energy efficient, high speed, and localized treatment solution for pathological disorders. However, most of the implantable devices developed so far stays on, or near the surface of the skin, as deep tissue implants introduces extra concerns in power delivery and data communication.

3.1.1 Power Delivery for Deep Tissue Implants

To deliver power wirelessly into deep tissue implants, the first concern that immediately arises is the path loss. Electromagnetic wave at GHz radio frequency usually faces a high attenuation, typically around 10 dB/cm [57, 58]. Yet ultrasound waves, commonly found in MHz frequency range, suffers from only $0.5 \sim 1.0 \text{ dB/cm-MHz}$ a loss in various tissue media. Together with a much higher permitted energy limit of 7.2 mW/mm$^2$ by FDA [59] (peripheral vessel only, 0.94 mW/mm$^2$ for abdominal region), ultrasound has emerged into a promising new power carrier for deep tissue implants.

This has motivated much research on deep tissue biomedical implant using ultrasound as the power and data telemetry modality [60, 61, 62, 63]. Yet to author’s best knowledge, all previously published ultrasound implants use ultrasound waves in a way similar to radio frequency electromagnetic waves. Such implants are designed to harvest focused continuous wave ultrasound, and information is transmitted using one of the classic data modulation methods, like FSK, ASK, or
BPSK, etc. Ultrasound, in this situation, is nothing more than just an alternative to electromagnetic wave as the physical layer media, where piezo elements replace the traditional antennas.

Figure 3.1: Three different cases for spatial power distribution when powering up deep tissue power harvesting implants.

However, such a direct translation can lead to a subtle problem when used with miniaturized, deep tissue, power harvesting implants. Considering two different scenarios shown in Figure 3.1 (A) and (B), where an implant has been placed deep beneath the skin. After the surgery site has healed, the exact location information of this implant is lost, but instead, a rough sense of where it might be is still known before the external device starts to interact with the implant. To power the implant up, the first solution, shown in Figure 3.1 (A), is to use the source transducer to broadcast its power over a large volume. This gives a high chance for the implant to pick up the remotely delivered power, and send back a response. However, it is obvious that all the spatial power are wasted except for the tiny fraction that directly covers the location of the implant, leading to a trade-off between the chances to find the implant, and the power delivery efficiency. To increase power transfer efficiency, focused power delivery can be used, as shown in Figure 3.1 (B). However, without the knowledge of the actual location of the implant, where should transducer focus its power to? In an application where only focused delivery can transmit enough power to generate a detectable response on the implant side, either a secondary mechanism is used to locate the device first, or a significant amount of work needs to be done on marking the location and blind searching...
for detectable responses.

To solve this problem, this work proposes the usage of a scanned focused wave, as shown in Figure 3.1 (C). If the device can harvest the energy in the period when the focused wave sweeps over to generate a detectable response, the location of the device can then be retrieved. Establishing a localized, focused power transfer now becomes possible. Yet this picture still has one important factor missing. Human beings, while alive, even when staying still, has unconscious movements like heart beating and breathing. These unconscious motion will move the implants in the surrounding tissue, not only in terms of the absolute position in space, but also the relative position to the human skin, a surface on top of which the external device is placed. This problem becomes more pronounced, when miniaturized implants are designed for minimal invasiveness, a key to long term implantation, as they have an intrinsically smaller reception area for incoming power. However, this is not a problem if the implant can function properly entirely under heavily duty cycled power delivery: as long as the sweep region covers the location of sensor, not only can the implant achieve its full function, but also can the user acquire the real time movement information of the sensor, a feature simple volume broadcasting cannot offer.

The ultrasound wave, used in an ultrasound sonography session, is exactly a kind of scanning focused wave.

3.1.2 Ultrasonography

Since pulse-echo mode ultrasound was first used in medical field [64], ultrasonography, or medical ultrasound, has emerged into one of the most widely used imaging procedure for real-time tomography. Before delving into the design details for the frontend circuit that is compatible with ultrasonography, it is important to understand the medical ultrasound’s working principle, and the typical considerations to optimize its performance.

Figure 3.2 depicts the working principle of the pulse-echo mode medical ultrasound. First, a focused, short ultrasound burst (commonly referred to as an ultrasound “beam”) is sent from the imaging transducer (Figure 3.2 (A)), forming a pulse that travels along a straight line into the
tissue. When this pulse hits the interface between two different tissue medium, a fraction of the power will be reflected from the acoustic impedance mismatch (defined as $Z = \rho c$, where $\rho$ is the density, and $c$ is the speed of the sound in the medium), forming an echo that travels back to the source transducer (Figure 3.2 (B)). The source transducer, after sent out the burst, has switched into receiving mode, and will record this reflected echo. By examining the time difference between the source pulse and all the received echoes, a one-dimensional line can be reconstructed if the speed of the sound in the medium is roughly known, reflecting the locations of acoustic impedance changes found along this line. By sweeping this focused beam from one side to the other (Figure 3.2 (C)), the acoustic impedance mismatch map can now be generated for a large two-dimensional cross section, called the field of view.

Several performance considerations become obvious once this simplified principle is introduced. The first involves the time duration of the source pulse. This directly affects the vertical resolution of the imaging process, as it directly translates to the width of the reflected echo. Hence shorter pulses are preferred for a higher resolution. The second involves the time difference between adjacent beams. To generate an image up to a depth of $z$, the wait time between adjacent beams needs to be at least $t = 2z/c$, as both the source pulse and the echo needs to travel along the medium. The third
consideration involves the spatial separation between each beams. If all beams can be formed as perfect lines, the spatial separation should be around the beam’s width $w$. However, in practice, perfect beam forming does not exist.

![Diagram of pulse-echo mode ultrasound imaging](image)

**Figure 3.3:** The working principle of pulse-echo mode ultrasound imaging.

Figure 3.3 shows a simulated focused beam. At far field, constructive interference happens. In this region, ultrasound energy is concentrated along a straight line with a width of $w$, and this is why it is usually called a “beam”. This desired beam is also called the “main lobe” in antenna design terminology. However, weaker “side lobes” also exists around the main lobe. These side lobes aliases the reconstructed image. Techniques to reduce side lobes’ intensity exists, for instance [65], but is considered beyond the scope of this thesis.

To design a frontend that is compatible with B-mode imaging, it is important to use the right testbench that reflects the nature of pulse-echo mode ultrasound received on the devices’ side, with anticipated non-ideal side effects that is unavoidable in medical imaging processes. And once proper functionality is achieved, the optimization of such circuits should also be done in a way that is in line with how medical ultrasound should be optimized, i.e., the circuit should be optimized to handle shorter pulses with weaker side lobes.
3.2 System Level Design

In this section, system level considerations are presented. Figure 3.4 shows the targeting use case, to demonstrate the concept of the designed ultrasound sonography compatible sensor system, and how it adds to a traditional ultrasonography session. Before the imaging session, one or more such devices are implanted deep underneath the skin. A traditional ultrasound B-Mode sonography imaging transducer can be then used to retrieve a live film of a cross-section of the inner body structure. As long as the devices are within the field of view of the imaging transducer, they will produce a mismatch of acoustic impedance relative to their surrounding tissues, shown as brighter spots in the reconstructed film. On top of this static ultrasound energy reflection, from time to time, the intensity of this reflection also changes, carrying the data representing what has been measured physiologically within the body. In this way, multiple implants can be distinguished from their different locations in the reconstructed image, with high level of confidence from their distinct data pattern that any passive devices cannot produce. To be fully compatible with imaging ultrasound, and to allow parallel data transfer at the same data rate for different implants, both data downlink (defined as from the transducer to the implant), and data uplink (defined as from the implant to the transducer) are synchronized to the frame rate of the reconstructed sonography movie.

At the transducer side, from frame to frame, the changes in intensity in the reflected echo are translated into bits. Thus from a network point of view, image reconstruction and spatial data detection can be seen as the mechanism implementing the physical layer of the data link. On top of this binary data, a protocol is required to ensure synchronized, interpretable data transmission, separating a continuous stream of data into frames, implementing the link layer of the data transfer. Finally, within the payload of the frames, an application implementation consist of instructions and data needs to be designed to dynamically control the behavior of the implants, and retrieve the physiological information, implementing the application layer of the data transfer. There is no network nor transport layer in the design, as the scenario does not allow peer-to-peer connections, and assume no congestion with perfect in-order data delivery.
With a high level picture of the desired behavior of such implants described, it is time to investigate how such intended behavior can be implemented on both the implant side and the transducer side.

### 3.2.1 Sonography Ultrasound Wave Seen from the Implant

The basic principle of B-Mode imaging was introduced back in Subsection 3.1.2. However, to design a device operating with the a sonography transducer array, it is more important to investigate what kind of ultrasound pattern the device will see, and how this is different from a traditional electromagnetic-like waveform used in communication circuits. For simplicity without generality, this chapter will only discuss B-mode imaging sessions using linear array transducers. The conclusion can be easily translated into cases with phased arrays, and it remains valid for most
pulse-echo mode imaging principles.

In an unrealistically ideal imaging scenario, a perfectly focused straight beam sweeps across the space from one side to the other. And the implant picks up only one strong pulse within each frame. However, such case is no even possible with ideal piezo elements in a homogeneous medium, as shown in Figure 3.3 in Subsection 3.1.2. For example, Figure 3.5 shows a simulation using an ideal linear array transducer. In Figure 3.5 (A), (B), and (C), focused beams scans from the left to the right, where the implant (gray box) is placed at the center. Aside of main lobes, the implant will also pick up ultrasound pulses at reduced magnitudes due to side lobes from beams targeting adjacent area, shown in Figure 3.5 (D). In analogy to the name of “main lobe” and “side lobe”, in this work, the strong pulse picked up from the main lobe is called the “main pulse”. And similarly, “side pulse” is used for ultrasound waves received at the implant from unwanted but unavoidable imperfection in the beam forming process. If the energy of all the pulses in one frames dies out before the next frame starts, the amplitudes in this pulse train is periodic, and repeats every frame. And here, all the pulses picked up in one frame is called the “pulse packet”.

Another important note is that the idea of having one or very few number of “main pulses” that has a much higher magnitude than other pulses in a pulse packet only holds well when the implant is placed in the far field of the transducer array, where coherent interference happens. In the near field before focusing is well established, the pulse packet will consist of a much larger number of pulses that is closer to the number of elements used to form the beam at a similar magnitude (see Figure 3.3). More generically, although an implant within the field of view will receive ultrasound energy in the format of pulse packets, the exact energy distribution within the pulse packet is a function of where the implant is relative to the imaging transducer. This is also one of the reasons why data transfer is designed to be synchronized to the frame rate: the pulse packet always has the same period as a frame, but if the downlink data changes in a time shorter than such period, the exact bit each device pick up in the full pulse packet will be a function of how exactly the beam is formed, how close the implants are to each other, as well as how close each implant is to the transducer, leading to a much more complicated data recovery and delivery scheme.
Figure 3.5: Ultrasound waveform received at the implant as a linear array transducer scans across the body.

One more important concern is that even if there are more than just one pulse hitting the implant for every frame, its power distribution in time is still extremely sparse. Considering the main pulse along: a 4 MHz, 4 cycle (which is 1 μs long) pulse received at the implant every frame, at 50 frames per second (fps) leads to a power duty cycle of 50 parts-per-million (ppm). Although side pulses exists, their magnitude are usually lower than the main pulse. Whether they contribute to a higher harvested power or not, and if so how much power can be harvested from them all remain in question and heavily depend on the actual topology of the rectifier (will be further discussed in Subsection 3.3.1). This provides challenges in system level specification.

3.2.2 Top-Level Modular Design

The idea of modular design is to partition the entire system into different function blocks, where the interfaces between these functions are somewhat “well-defined”. The reason for this word under quote is that in analog designs, blocks’ performance can be rarely well isolated as all block loads
each other in one way or another. Although the level of “well-defined-ness” found in analog circuits are usually low, finding interfaces where loading effects are relatively easy to understand and control with a reasonable amount of over design in presence still polishes designer’s understanding to the problem, and guide designer’s energy into more specific optimization processes.

Figure 3.6: Block diagram for the sonography compatible implant.

As illustrated in Figure 3.4, the implant side can be divided into 3 parts according to their functionality in the data protocol stack, i.e., the physical layer, the link layer, and the application layer. This is illustrated with more details in Figure 3.6. The physical layer implements bit-to-bit data communication between the implants and the transducer. The link layer group bits together into frames for interpretation. Application layer defines the behavior supported by the implant, with a system-specific ID implemented, and accessible through instructions to allow pseudo-parallel, non-interfering, device specific instruction execution in a distributed implanted sensor network scenario.

Once physical data can be interpreted reliably, the link layer and the application layer’s functionality are mostly logic driven where existing solutions can be re-purposed for this specific application. However, for an energy harvesting implant, the establishment of the physical layer connection also requires reliable and sufficient energy harvested from the source. This is even more
problematic as the power incoming are deeply duty-cycled. Thus physical layer implementation poses unique challenges, which will be addressed in detail later in Section 3.3.

On the top level, the design of the physical layer is targeting a power delivery of 100 pW from 4 MHz, 50 ppm pulsed ultrasound to the digital logic (link layer and application layer), while the digital logic is targeting at a clock speed synchronized to the frame rate of the imaging session (assuming 50 fps), consuming less than 100 pW. The link layer frames are set to 16 bits long. Active data transfer is difficult with such a low power available, thus modulated backscatter is chosen to implement the data uplink. By modulating the electrical impedance loading the piezo element, the acoustic impedance of the piezo crystal will also change through its intrinsic electrical-mechanical coupling. This will in turn change the intensity of the reflected echo when an ultrasound pulse reflects at its surface. In this prototype, the amplitude of the reflected backscatter intensity change is chosen to be maximized, which leads to a completely short of the piezo when a bit “0” is to be transmitted. This design choice complicates the implementation of the data uplink, which will be documented in Subsection 3.3.5

3.2.3 Modification to the Existing Imaging Process

Although the idea is to build an sensor interface that is as compatible to the native imaging process as possible, certain modification is still required, as there is no built-in mechanisms for the ultrasound transducer to send data down into tissue. Data downlink in this version is implemented using pulse-width-modulation (PWM). For a data “0” sent down from the transducer, a shorter pulse is used in every beam for the same frame. For a data “1”, a longer pulse is used.

This solution is certainly not optimal, as a varying pulse width from frame to frame creates a changing vertical resolution, and thus a “flashy” movie. This is resolved by not reconstructing the image when data “1” is sent for that frame, converting the “flashy” experience into a “laggy” one.
3.2.4 Ultrasound System

To implement the modification mentioned in Subsection 3.2.3, the ultrasound system needs to allow easy access to modify existing imaging processes. In this design, the Verasonics Vantage 256 research ultrasound system is chosen with L12-3V linear array transducer (Verasonics Inc., Redmond, WA, USA) as the B-mode imaging system for all the experiments. The choice of this specific ultrasonography system has led to easy prototyping through its MATLAB script interface (Mathworks Inc.). But certain built-in characteristics in this system may in turn generate sub-optimal design choices on the implant side.

One of such design choices taken driven by the specific implementation of the ultrasound system, is the choice of operating frequency. The Verasonics system is fully digitally controlled. This includes the operating frequency as well as the driving waveforms on the piezo transducers. The center frequency of the ultrasound wave is down counted from a master clock, leading to a specific, discrete subset of available frequencies that can be used for testing purposes. The driving waveforms use discrete 3-level digital pulses, leading to nonlinearities and long tails in the generated acoustic waves. These limitations has led to certain non-optimal design considerations that will be introduced later in Subsection 3.3.1, though such design choices have, in turn, gave the implemented system a potentially much wider robustness in real-world application.

Another factor that will drive future design choices is the exact formation of the imaging process. In this design, minimum modification to the default imaging script is used as a baseline for the prototype design. And for all the testing, 192 focused, 4 MHz center frequency ultrasound beams, or ray-lines scan linearly to form one 2-dimensional cross section image, with 100 µs delay between adjacent ray-lines. The scanning happens at 50 fps.

3.3 Circuit Design

In this section, the detailed transistor level implementation of each block is documented. The majority of the section focuses on the implementation of the physical layer (see Subsection 3.2.2),
as it is the most challenging functionality to realize. It includes a switched-only rectifier and a voltage regulator for stable supply generation, clock and data recovery circuits, an uplink modulator for bi-directional data transmission, and a minimal link layer and application layer protocol aiding the performance verification.

3.3.1 Switch-Only Rectifier

To harvest ultrasound energy, a rectifier is needed to convert the ac electrical waveform generated by mechanical excitation received on the piezo element into a dc voltage to power up the rest of the system. The design of the rectifier is dictated by the equivalent electrical impedance looking into the port of the piezo element, and the properties of the incoming electrical waveforms. As a transducer between mechanical domain and electrical domain, the electrical impedance of a piezo is also a function of its mechanical boundary condition.

An in-house measured electrical impedance (in Z-parameters) is shown in Figure 3.7. Figure 3.7 (A) and (B) shows the impedance of a 1 mm wide by 1 mm long by 0.5 mm thick lead zirconate titanate (PZT) from 300 kHz to 10 MHz, with (A) being the case of the PZT submerged in water (tissue phantom), and (B) being the case in castor oil (fat phantom). It can be found that over the tested spectrum, the impedance of PZT remains mainly capacitive, with no resistive series resonance found. Although other works reported inductive impedance observed in piezo crystals that can be used for high efficiency power transfer [66, 67], yet the author finds that the existence of inductive impedance, the frequency range inductive impedance can be found, and the equivalent inductor value all heavily depend on the mechanical properties directly loading the piezo element. For instance, in Figure 3.7 (C) and (D), the impedance is measured again under either water (C) and castor oil (D), but with a lead magnesium niobate-lead titanate (PMN-PT) crystal with the same size at 0.48 mm thickness, together with an air pocket physically attached to one side of the piezo. With the addition of the air pocket, series mode resonance is now found in both cases. However, this was not carefully examined at the time of design, and how the impedance of the piezo element changes with biological tissue’s mechanical loading is still not well investigated. Also, it
is still a question whether the discrete set of frequency supported by the ultrasound system in use (see Subsection 3.2.4) can generate ultrasound pulses that falls into the inductive range, and how accurate the mechanical fabrication flow can be controlled.

In this design, the conservative approach was taken, which assumes the piezo element always shows a capacitive impedance across the frequency of interest. For the highest energy efficiency, conjugate matching is certainly desired, however, in this case, to cancel the reactance, an inductor at 10s of $\mu$H is still needed for PZT across the frequency of interest (Figure 3.7 (A) and (B)), significantly increases the overall size of the system. This problem is not new to the field of piezoelectric energy harvesters. For low frequency shock energy harvesters, the matching inductor
required is commonly at 10s of Henry level. With such a gigantic inductor required, matched power transfer is just impossible, leading to the definition of the “ideal full bridge rectifier efficiency”. It can be proved (for instance in [68]) that for an ideal full bridge rectifier, the maximum power this rectifier can harvest from a capacitive voltage source is:

\[ P_{FBR} = C_P V_P^2 f_P, \]  

(3.1)

where \( C_P \) is the source capacitance, \( V_P \) is the voltage amplitude when the piezo element is in open circuit, and \( f_P \) is the frequency of excitation. However, modern piezo harvester design has advanced to a state where through the usage of switch-mode impedance matching, efficiencies around 30% can be achieved at sub-100 Hz input excitation frequency range [69, 70]. This has inspired the author to translate one of such techniques, known as “switch-only”, to a higher frequency range, and see how well it boosts the energy harvesting efficiency. Techniques with even higher efficiency typically requires mH-level inductors and/or nF-level capacitor array, which will significantly increase the overall implant size. These techniques are not adopted to maintain an implant friendly size of the envisioned system. The schematic of the proposed switch-only rectifier is shown in Figure 3.8.

![Schematic of the switch-only rectifier.](image)

Figure 3.8: Schematic of the switch-only rectifier.

A conceptual waveform is shown in Figure 3.9 to better illustrate the working principle of
the switch-only rectifier. The key efficiency enhancement comes from the observation of the “conduction angle” of the full bridge rectifier. After $V_{PZ+}$ becomes just lower than the power rail $V_{CC}$, the full bridge rectifier stops conducting. The full bridge rectifier now waits the electrical signal on the piezo element enters the next conduction phase, i.e., $V_{PZ-}$ becomes higher than $V_{CC}$, causing a long “off-time” in which no power is harvested. To reduce the off-time when energy is not harvested on the storage capacitor, a switch can be added in parallel with the piezo crystal that shorts the two terminals of the piezo for a brief period of time right after the piezo exits the conduction angle. For an ideal switch, this will zero the voltage across the piezo, effectively accelerating the reverse charging of the input capacitor, such that less time is required to charge $V_{PZ-}$ to a level higher than $V_{CC}$, effectively increase the conduction angle. Through this “switch only” technique, a maximum amount of $2 \times P_{FBR}$ power can be harvested [68].

![Conceptual waveforms for the switch only rectifier.](image)

**Figure 3.9:** Conceptual waveforms for the switch only rectifier.

On top of a likely-capacitive input impedance, there is one more challenge that this rectifier needs to address. As discussed previously in Subsection 3.2.1, the energy received during an ultrasound imaging session is in the form of heavily duty-cycled pulses. This means the power harvesting only happens at a very short of period of time, but the leakage from the rectifier will
discharge the storage capacitor all the time. The performance of a passive rectifier is then directly limited by the on-off ratio of devices provided by the technology. An active rectifier, though possible to surpass this limit, poses another challenge in the implementation of the active diode, or more specifically, the comparator in the active diode.

Figure 3.10: Schematic of the comparator used in the active diode.

The proposed comparator is shown in Figure 3.10. If the incoming wave is a continuous one, then the gate bias voltage of $M_{N1}$ and $M_{N2}$ can be tuned to find the comparator’s best trade-off point between the speed and power consumption, such that the harvested power is maximized. For a heavily duty-cycled input, to harvest the 4 MHz ultrasound pulses that come in every 20 ms, the bias voltage needs to be low enough to reduce the leakage from $V_+$ (connected to $V_{CC}$) through $M_{P2}$ and $M_{N2}$ to ground, but also high enough to generate the output decision with a negligible delay in a tiny fraction of the 250 ns period. Such a trade-off point does not exist using devices provided in the technology. To solve this problem, a dynamic biasing topology is implemented, where the incoming wave after the full bridge rectifier is directly used as the bias voltage for the two current source transistors, breaking the trade-off between static power consumption and dynamic speed. In this case, when there is ultrasound pulses present, the speed of the comparator increases as the
sine wave inches higher, and become fast enough for a decision when $V_{FBR}$ is comparable to $V_{CC}$, to turn on or off the active diode.

However, this design choice certainly creates a current consumption overhead when $V_{FBR}$ becomes too high. Several voltage limiting techniques was tried out at the design time, yet all of them introduce potential power up deadlocks, thus were not adopted. The author believes techniques improves the performance of the comparator exists.

With this comparator topology, it is now a good time to investigate how the existence of side pulses will affect energy harvesting. At cold start up, where $V_{CC}$ is around 0 V, the amplitude of side pulses is higher than $V_{CC}$, and thus they will contribute to a higher energy collected at the storage capacitor. However, when $V_{CC}$ is well established, a side pulse that is slightly lower in amplitude than that of the $V_{CC}$ will introduce a high, transient, “cross-bar”-current-like leakage from $V_{CC}$ to ground through $M_{P2}$ and $M_{N2}$, thus draining the stored charges without charging back. This will ultimately limit how high $V_{CC}$ can reach. To summarize, with this version of the rectifier, side pulses help the cold start up process, but they are likely to become a power limiting factor for how high $V_{CC}$ can get to.

The target voltage at node $V_{CC}$ is 1.2 V, yet this value is by no means regulated. All circuits powered at this domain needs to function properly when $V_{CC}$ is higher than 1.2 V, but should be designed in a way less problematic when $V_{CC}$ is too high (up to 1.8 V). A storage capacitor of 100 pF is chosen as a trade-off between power line noise and the start up time.

3.3.2 Voltage Regulator

To achieve a better defined behavior for the majority of the circuit, a voltage regulator is required to generate a known voltage. The reference voltage generation is implemented using a PMOS-only, trim-free structure, a technique first introduced in [71]. Although designs like [72] can achieve an even lower power consumption at a reduced area, it has an intrinsic sensitivity to process variation from the usage of different types of MOSFETs, requiring calibration and trimming. The PMOS-only structure, on the other hand, can be well matched, and can be modified slightly to trade its
temperature sensitivity for a much lower power consumption. The simplified schematic of the voltage reference is shown in Figure 3.11.

![Schematic of the voltage reference.](image)

Figure 3.11: Schematic of the voltage reference.

The working principle of this voltage reference can be derived as follows. For the $M_1$ and $M_2$ branch, assuming deep-subthreshold saturation ($V_{DS}$ is dropped out from the equation):

\[
I_{D,M1} = I_{D,M2}
\]

\[
I_S \frac{W_1}{L_1} e^{(0-V_{TH,1})/n\phi_T} = I_S \frac{W_2}{L_2} e^{(0-V_{REF}-V_{TH,2})/n\phi_T}
\]

\[
V_{REF} = (V_{TH,1} - V_{TH,2}) + n\phi_T \ln \left( \frac{W_1/L_1}{W_2/L_2} \right)
\]

(3.2)

A similar calculation can be done for the branch consists of $M_3$ and $M_4$. Since $M_3$ and $M_4$ both have $V_{BS} = 0$ V, $V_{TH,3} = V_{TH,4}$, the above equation can be simplified to:

\[
V_{DD} - V_{B,1} = n\phi_T \ln \left( \frac{W_4/L_4}{W_3/L_3} \right)
\]

(3.3)
Finally notice that body effect follows:

\[ V_{TH,1} = V_{TH,2} + \gamma \left( \sqrt{2\phi_F - (V_{DD} - V_{B,1})} - \sqrt{2\phi_F} \right) \]  \hspace{1cm} (3.4)

The reference voltage can be expressed as:

\[ V_{REF} = \gamma \left( \sqrt{2\phi_F - n\phi_T \ln \left( \frac{W_4/L_4}{W_3/L_3} \right)} \right) + n\phi_T \ln \left( \frac{W_1/L_1}{W_2/L_2} \right) \]  \hspace{1cm} (3.5)

In the Equation 3.5, the first term is complementary to absolute temperature (CTAT), while the second one is proportional to absolute temperature (PTAT). Thus by adjusting the ratio of the transistor sizes, a possible low temperature dependency at the output \( V_{REF} \) can be achieved. Notice also, that the cancellation between the PTAT term and the NTAT term are not mathematically matched, thus these values, corresponding to the two separate branches (the \( M_1 M_2 \) branch and the \( M_3 M_4 \) branch) can be implemented using different types of transistors. If only a narrow temperature compensation range is required, then the drain current in the design can be further drastically reduced, as a single point of cancellation can be chosen without worrying about its behavior far from the compensation point. This is indeed the case in this design, as the author only cares about room temperature to high body temperature (20 °C to 42 °C). By using I/O devices in the \( V_{REF} \) branch, the final expression for \( V_{REF} \) is:

\[ V_{REF} = \gamma \left( \sqrt{2\phi_F - n_{Core}\phi_T \ln \left( \frac{W_4/L_4}{W_3/L_3} \right)} \right) + n_{I/O}\phi_T \ln \left( \frac{W_1/L_1}{W_2/L_2} \right) \]  \hspace{1cm} (3.6)

Keen readers may find that the dependency of \( I_S \) on \( V_{BS} \) (see Equation 2.2) is omitted in the derivation. Yet since the design process requires a heavy empirical optimization (or, “Spice-monkey-ing”), second order dependencies can be resolved in a numerical iteration fashion.

In this design, less than 22 pW power consumption is achieved up to a \( V_{CC} \) of 1.8 V, while only 9.4 pW is used at desired \( V_{CC} \) of 1.2 V at a nominal temperature of 37 °C. The output voltage is 304 mV, with about 0.6% variation across the temperature range of interest. Capacitors not shown
in the schematic are added to enhance the PSRR to the output reference voltage.

![Figure 3.12: Schematic of the voltage regulator.](image)

With a voltage reference designed, it is now possible to design the complete voltage regulator. The schematic of the voltage regulator is shown in Figure 3.12, which adopts a low-dropout (LDO) like topology, but replaced the PMOS power transistor with a native NMOS transistor. Since the output current is below 100 pA, the gate voltage $V_G$ needed for the NMOS transistor is kept well below $V_{CC}$, and within the output range of the OTA. The usage of an NMOS pass transistor significantly boosts the PSRR of the regulator, at a ratio of approximately $g_m/g_{DS}$. This is especially important, as a lot of power noise is expected in $V_{CC}$ at 50 Hz (the frame rate) and 4 MHz (the center frequency of the ultrasound). The resistor ladder, implemented by a series of 10 diode-connected PMOS transistors, sets the output voltage at $V_{DD}$ to be roughly 506 mV in simulation.

The schematic of the OTA is shown in Figure 3.12 (B). PMOS side cascode is used to reduce systematic mismatch, as the output voltage is usually low.

Now that the voltage regulator is in place, a well defined $V_{DD}$ is ready to be used across the system.
3.3.3 Clock Recovery

The key functionality of the clock recovery circuit is to generate the on chip clock signal that is synchronized to the frame rate of the imaging process. The schematic and the designed behavior is shown in Figure 3.13 (A).

![Schematic and waveform of the clock recovery circuit.](image)

Here, an S-R latch (input S and R are inverted) is used to detect the incoming pulse harvested at the piezo element, and produces the falling edge of the output clock. Signal $\bar{Q}$ is also set to “0”, and after a “blackout” delay $\tau_{CLK}$, resets the S-R latch, which in turn triggers the rising edge of the clock. By design, side pulses comes in during the “blackout” delay will not trigger more clock flips. Since the first pulse that is higher than the transition voltage $V_M$ of the inverter will trigger the clock, it is important to have this delay $\tau_{CLK}$ longer than the period where strong side pulses may occur, to avoid these side pulses corrupting clock synchronization. Assuming 3 strong side pulses in far field (7 pulses total that reaches $V_M$ in one pulse packet), and other timing assumptions from Subsection 3.2.4 (100 $\mu$s between adjacent pulses, and 50 fps), the requirement for this delay is $600 \mu$s < $\tau_{CLK}$ < 20 ms. This design margin is wide enough to implement delays reliably using devices in subthreshold, which suffers from an intrinsically higher sensitivity to process, voltage and temperature (PVT) variations. Another important characteristic of the blackout delay generator
is that only the falling edge delay is necessary for a proper functionality. In this work, the time constant is generated from a gate-source connected PMOS’s subthreshold leakage current charging a capacitor $C$ (shown in Figure 3.13 (B)), taking advantage of the wide design margin for single edge sensitive delay generation.

Regulated $V_{DD}$ is used only in the blackout delay generator to reduce potential variations in this delay value, with level shifters to drive the reset input of the S-R latch. Other blocks are power at the unregulated $V_{CC}$. The distribution of the clock signal CLK at the global scope is also at $V_{CC}$ level, as a much higher overdrive voltage gives negligible delay in the chip level clock propagation, which in turn minimizes the efforts (area and power overhead) in timing closure (setup and hold across PVT) during logic synthesis. Driving several 10s of fF load, this provides an acceptable power overhead in the global clock distribution.

Finally, the rising edge of the clock is chosen to be the one other than what is triggered directly by the incoming pulse. This gives a much more relaxed timing constraint, as digital logic need to determine the uplink data bit after the entire pulse packet is over, such that all the beams (it is unknown to the IC which one is the “main beam”) picks the same data bit, but before the next pulse packet comes in.

3.3.4 Downlink Data Recovery

To recover the frame-to-frame downlink data bit embedded in the width of each received pulse, the downlink data recovery circuit is designed to count the number of cycles found in one pulse (schematic and waveform shown in Figure 3.14). Although the pulse being counted can be any pulse in the same pulse packet, it is natural to choose the one that also triggers the clock’s falling edge (see Subsection 3.3.3). A gating signal “EN” is generated to block the ultrasound signal feeding into the counter after the first pulse, with the help of a delay element. In this case, the delay element is implemented using a 3-stage inverter chain. Its delay value needs to fall into the range of $2 \mu s < \tau_{DATA} < 100 \mu s$, assuming a maximum 8 cycles is allowed. This margin is, again, wide enough for such a topology. The counter resets after each clock rising edge, and is powered at
unregulated $V_{CC}$ to keep up with MHz-level ultrasound signal. The decision logic, however, only needs to operate once per frame, thus is connected to $V_{DD}$ for a reduced power consumption.

A brief outline of the behavior of the decision logic is written in pseudo code as follows:

Input:  reset, count, clock  
Output:  data_in, data_ready  
Reg:  value_threshold, value_max, data_prev, confuse_counter  
Param:  confuse_limit = 32

Reset:  

```
value_threshold <= 0, value_max <= 0, data_prev <= 0, 
data_in <= 0, data_ready <= 0, confuse_counter <= 0
```

Operation (synchronized to clock rising edge only):

```
data = (count < value_threshold) ? 0 : 1  
if (count > value_max):  
    value_max <= count  
if (count < value_max - 2):  
    value_threshold <= count + 3  
data_ready <= 1  
data_prev <= data
```
if (data_prev == data):
    confuse_counter += 1
if (confuse_counter == confuse_limit):
    goto reset

Listing 3.1: Pseudo code for downlink data recovery decision logic.

The confuse counter in the pseudo code implemented in the decision logic gives a safe guard if a random interference is picked up, corrupting the history-based data recovery scheme. With the confuse counter, the data recovery logic resets itself if no different data is detected within a limited amount of cycles. The existence of the confuse counter also makes certain type of data pattern more welcome than others, which will be discussed later in Subsection 3.3.7.

3.3.5 Uplink Data Modulator

The uplink data modulator is probably one of the most complicated circuit within the designed system, as it can be hardly isolated, and its functionality couples with the power harvesting as well as the synchronized logic portion of the system. Hopefully this subsection can offer a comprehensive overview of the design choices taken to implement a reliable uplink data transmission. The majority of the complication comes from the fact that backscatter is used as a low power method to fulfill this data uplink.

The schematic of the uplink data modulator is shown in Figure 3.15. To optimize the data modulation depth in the backscatter (see Subsection 3.2.2), a completely “on/off” modulation is taken on the electrical side of the system, i.e., the NMOS $M_{NB}$ used for this backscatter modulation will see a digital level between the on chip ground $V_{SS}$ and highest unregulated supply $V_{CC}$. This “on/off” level will only translate to an amplitude shift in the acoustic impedance $Z_P$ of the piezo element, and thus only modulates the backscatter echo with a limited depth through the change of the reflection factor, though this is already the highest modulation depth the system can produce. But this implementation gives rise to several problems. The first one is that now when a “0” (switch $M_{NB}$ turned “on”) is to be transmitted, the piezo appears to be shorted for the rectifier, thus no
power can be harvested. This leads a reduced amount of power available to operate the circuit, and is resolved by power budgeting in the top level. The interface power budget discussed in Subsection 3.2.2 already has this taken into consideration. The second one is that when the piezo is shorted, no clock recovery happens, and thus stalls the entire synchronized logic. This is shown in detail Figure 3.16.

After the first clock rising edge where an data “0” is decided to be sent (“D_UP”, stands for “DATA_UP”, which is not shown in the schematic in Figure 3.15), no further clock signals can be recovered. However, the value of “DATA_UP” is generated by the synchronized logic, meaning without a rising clock edge, it is also stuck at “0”. This is effectively a deadlock, in which before something else triggers an effective reset behavior, the state of system will not change anymore. To solve this, the uplink modulator needs to bring a initial “0” back to a “1” in an asynchronous fashion. This is implemented using a 5-bit tunable falling edge delay element and a backscatter tuner logic. Upon starting up, the uplink data transmission is blocked, and a linear search is performed to find a suitable delay value. This delay value needs to be as close to one clock period as possible. The schematic of the 5-bit tunable delay element is shown Figure 3.17.

However, within above approaches, one more problem now becomes apparent. Instead of
Figure 3.16: Illustration of the case when uplink corrupts the clock recovery without a proper modulator.

resetting “DATA_UP”, it instead resets the backscatter node “BS”. “0”’s that follow up the initial “0” will not trigger the delay elements anymore, and output remains at value “1”. To solve this, on top of a main data path “DATA_M”, a matched auxiliary data path “DATA_A” is required to handle consecutive “0” transmission. When “0” needs to be transmitted in adjacent bits, the main path and the auxiliary path will flip alternatively, to ensure a correct output.

Finally, one minor problem this brings is that when a “0” is to be transmitted, one and exactly one clock cycle will still be missing, causing a bit “0” taking two frames in the reconstructed movie to transmit. To implement a constant data rate in the uplink, all “1”’s are also extended by one clock cycle, such that in a 50 fps session, the uplink data rate is a stable 25 bits/s.

To sum up, a pseudo code description of the behavior of the backscatter tuner logic is available as follows, with waveforms shown in Figure 3.18.

Input: reset, clock, bm, data_up
Output: delay, en, data_m, data_a
State: TUNE, READY

Reset:
  delay <= 31, en <= 0, data_m <= 1, data_a <= 1
Figure 3.17: Schematic of the 5-bit tunable falling edge delay element.

Operation:

TUNE:

data_m <= 0

at next clock falling edge:

if (bm == 0):
    confirm at off

at next clock rising edge:

if (bm == 1):
    confirm at on

if (confirmed at both on and off):

    en <= 1
    goto READY

else:

    delay -= 1
    data_m <= 1
    goto TUNE

READY (synchronized to clock rising edge):

if (data_up == 0):

    if (data_m == 0):
        data_m <= 1
```python
    data_a <= 0
else:
    data_m <= 0
    data_a <= 1
else:
    data_m <= 1
    data_a <= 1
wait for one clock cycle
```

Listing 3.2: Pseudo code for uplink tuner logic.

Another minor problems that arises from shorting the piezo, is that when an uplink bit of "0" needs to be transmitted, no data downlink can be detected either. Thus data uplink and data downlink cannot happen simultaneously.

![Waveforms of the key signals in the uplink data modulator.](image)

The implementation of the uplink modulator, in order to maximize data modulation depth, has introduced scattered design choices across different seemingly uncorrelated function blocks. Yet the problem it is trying to solve is still clear: to deliver reliable uplink data at a constant rate, while minimizing disruptions to the other parts of the system.

3.3.6 Periphery Circuit Blocks

There are three important periphery circuit blocks that are key to the proper functionality of the system. They are the power-on-reset (POR), the fuse based ID generator, and a high power gateway.
Figure 3.19: Schematic for the power-on-reset.

For a remote standalone system, a known initial state is crucial for it to operate correctly, especially for the digital logic within it right after power up. Thus a POR is added into the system. The POR generates a reset signal, and only de-assert it after the entire circuit is configured to a known reset state. The schematic of the POR circuit is shown in Figure 3.19. The circuit uses a similar principle to all the delay blocks introduced previously, where the subthreshold leakage of a gate-source connected PMOS is used to charge a capacitor, generating a reset time $t_{POR}$ before the output signal of the POR goes high. To ensure proper functionality, the choice of $t_{POR}$ needs to be longer than the time taken for $V_{DD}$ to be properly regulated, plus the reset pin hold time required in the asynchronous reset/set flip-flops in the digital circuit.

Another important block is the fuse based ID generator, as it is the key to tell different implants apart during a distributed, pseudo-parallel operation. The schematic of the 1-bit ID generator is shown in Figure 3.20. Without any modification, the output will be a “1”, at the expense of the current consumption from one gate-source connected PMOS. If the exposed Metal 6 is etched, the output will now be a “0”. This simple implementation allows a pre-defined ID for each implant by physically removing one or more fuses. An 8-bit array of the fuses is used for an 8-bit on chip deterministic ID.

The final block is called the high power gateway. This is to implement a secondary functionality,
anticipating the application scenario in which a higher amount of power is required to operate the sensor that is integrated with the ultrasound frontend circuit. Once the location of the implant is known during the ultrasonography session through the identification of the implant’s data signature, focused power transfer becomes possible if the implant’s location is not subject to a significant amount of movement during the time required. Thus, a high power gateway is designed to switch into high power mode when specific instructions are sent (see Subsection 3.3.7).

After receiving the specific instruction “HP MODE”, the high power gateway will inspect each adjacent pulses, and see if there is a consecutive pulse width change between them. Since side pulses are unavoidable in a regular imaging session, and for frame-synchronized data delivery, all side pulses carry the same bit of information, this behavior will only happen when all the beam are directing at the implant, thus delivering a much higher amount of power (or more specifically, roughly the number of beams per frame times higher). At the same time, the high power gateway will block clock and data recovery, and delivering the power, a high speed data recovery, and the control of the backscatter switch to a potentially integrated sensor system.

Figure 3.20: Schematic for one bit fuse based ID generator.
3.3.7 Link Layer and Application Layer

To demonstrate the functionality of the proposed implant, a minimum effort link layer protocol is implemented in the chip, to support 6 application layer instructions.

The link layer frame structure for downlink data and uplink data are shown in Table 3.1. The starting “0101” in the downlink serves as a header, while an alternating code minimizes the chances for the device getting into “confused” state (see Subsection 3.3.4) when valid data is being transferred. The starting “00” in the uplink frame also serves as a header, however, since every bit takes two image frames, on the received side, this also looks like “0101”. The “P” at the end of an uplink frame stands for parity check (XOR, or one bit addition of all the bits in the frame), where both the parity value and its inverse are used to expand the frame into 16 bits long.

<table>
<thead>
<tr>
<th>Bit</th>
<th>15</th>
<th>14</th>
<th>13</th>
<th>12</th>
<th>11</th>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td>0101</td>
<td>INST</td>
<td>ARG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frame Structure for Data Uplink

<table>
<thead>
<tr>
<th>Bit</th>
<th>15</th>
<th>14</th>
<th>13</th>
<th>12</th>
<th>11</th>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td>00</td>
<td>INST</td>
<td>RET</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Frame structure for link layer.

The list of valid instructions (“INST”), arguments (“ARG”), and return values (“RET”) are shown in Table 3.2 to implement different application layer functionalities.

<table>
<thead>
<tr>
<th>Name</th>
<th>INST</th>
<th>ARG</th>
<th>RET</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUERY ID</td>
<td>0101</td>
<td>Do not care</td>
<td>ID from fuse array</td>
</tr>
<tr>
<td>HELLO</td>
<td>0110</td>
<td>ID</td>
<td>HELLO = 0101 0101</td>
</tr>
<tr>
<td>CONFIG</td>
<td>1010</td>
<td>ID</td>
<td>ACK = 1010 1010</td>
</tr>
<tr>
<td>STORE</td>
<td>1001</td>
<td>DATA</td>
<td>ACK = 1010 1010</td>
</tr>
<tr>
<td>LOAD</td>
<td>1100</td>
<td>ID</td>
<td>DATA</td>
</tr>
<tr>
<td>HP MODE</td>
<td>1101</td>
<td>ID</td>
<td>ACK = 1010 1010</td>
</tr>
</tbody>
</table>

Table 3.2: Application layer instructions.
It is worth noting that “0101 0101 0101 0101” is a valid instruction for “QUERY ID”, even if the first 8 bits went undetected. It is useful to apply this instruction, and measure the delay of the chip’s response in number of frames, to debug the data recovery circuit. Instruction “HELLO” is the simplest ID-specific instruction, and can be used to debug parallel, distributed operation. “CONFIG” and “STORE” instructions are in place to save specific values into the on chip register on the fly, a function for static configuration of the sensor interface circuits that will be integrated in the future. “LOAD” instruction will fetch the previously stored value back.

The link layer and application layer logic are synthesized using the digital standard cell library described in Subsection 2.4.1.

### 3.4 Performance Verification

The proposed IC (within the chip set USTAG V2.71) was fabricated in TSMC 180 nm 1P6M Salicide with AlCu FSG process (MSRFGII). A die photo is shown in Figure 3.21. The IC occupies an area of 830 µm × 740 µm with pad ring and ESD structures included. In the die photo, several blocks are labeled. The switch-only rectifier is found in the area surrounded by the red line, where the downlink data recovery as well as the uplink data modulator are placed in the blue boundary. The pink-purple boundary specifies the area taken by the voltage regulator. Finally the yellow boundary surrounds all digital blocks, including the implementation of link and application layer.

#### 3.4.1 Post Fabrication

The main goal of the post fabrication is to define an ID for each chip. This is done through chemically etching one or more exposed fuses in the 8-bit fuse array (see Subsection 3.3.6). The entire chip set USTAG 2.71 is first subject to standard photolithography process, through which a positive resist (AZ1813, MicroChemicals) based pattern is defined with the only the desired fuses exposed. The IC is then submerged in ferrochloride solution (FeCl$_3$, usually marketed and used as a copper etchant, but it also etches aluminum) for 20 minutes for bulk fuse removal. The top metal removal will reveal the seed layer at the bottom and the side walls of the fuse, between the foundry
Metal 6 and the passivation material. This seed layer can be further removed by submerging the chip in a mixture of 1 portion of 30% ammonium hydroxide (NH₄OH, Sigma-Aldrich) and 2 portions of 30% hydrogen peroxide (H₂O₂, Fisher Scientific), in 1 minute intervals with visual checks under the microscope in between. Note that this solution is highly unstable, and expires in around 10 minutes after mixing. It is also very reactive with many side reactions possibly unknown to the author. It is important to ensure a good quality in the photoresist layer, as well as to make periodic verifications of the intactness of the resist layer during the process. An example of one IC with the least significant bit etched is shown in Figure 3.22.

The poly-silicon layer stripe is added in the design time as a visual mark for a successful fuse removal, and numbers on the same layer are added to indicate which bit this fuse corresponds to. In the example shown in Figure 3.22, before etching (Figure 3.22 (A)), the IC has an ID of 0xFF. And after etching, the chip is assigned with an ID of 0xFE (Figure 3.22 (B)).

After the ID is specified to the IC through the post processing described above, the IC can now be packaged into a fully integrated and implantable device. The implantable device occupies 11
mm\(^3\) in volume, with an Rogers 4003C board hosting the IC, a PZT crystal with a size of 1 mm by 1 mm by 0.5 mm (PZT 5A, Piezo Systems), and two 0201 capacitors (Murata Manufacturing Co., Ltd) at 100 pF each for power line decoupling. A photo of the implant is available as Figure 3.23. The IC is thermosonically flip chip bonded in a process similar to what was described in [73], with the bonding process done using Fineplacer (Finetech) at 340 °C for 10 s on the package side, and 260 °C for 20 s on the chip side, under 4 N forces compressing the IC down against the electroless nickel electroless palladium immersion gold (ENEPIG) finished package board, with gold balls pre-attached using a wirebonder on all pads.

3.4.2 Electrical Measurement

The electrical testing is performed using wirebonded chip with a minimum functionality PCB, which incorporates an Opal Kelly XEM 6010 FPGA (Opal Kelly) that is controlled by a command line script to trigger the function generator (Figure 3.24), generating the emulated ultrasound wave to feed into the IC. A series 10 pF capacitor is added to emulate the source impedance of the piezo crystal. Test signals are wired to an oscilloscope for debugging purposes. All application layer instructions are first tested after a successful power up. They behave exactly as how they are designed, an indication of a potentially timing violation free and functionally correct digital logic
Figure 3.23: A photo of the fully packaged ultrasound sonography compatible implant.

The DC power up testing is performed using the probe station. With a linear slow ramp from 0 V to 1.8 V sent to pin $V_{CC}$, the regulated power supply $V_{DD}$ is probed as a measure of the DC power regulation characteristic. A relatively steady $V_{DD} = 530$ mV DC output is found at $V_{CC}$ between 1.2 V and 1.8 V, the designed working range, as shown in Figure 3.25. At the minimum $V_{CC}$ of 1.2 V, the IC consumes a static power of only 57 pW, thanks to the low leakage custom standard cell library, and the deep-subthreshold analog design enabled ultra-low power voltage regulator.

Power efficiency of the rectifier is probed using the setup described in Figure 3.24, with the duty cycle of the ultrasound pulses sweeping from 10 ppm to 1 (continuous wave). A standalone rectifier
implementation is used for the measurement, which is available in the chip set USTAG V2.71, and is wirebonded for this test. A transimpedance amplifier (TIA) is connected to the output of the rectifier, biasing its output voltage at 1.2 V while power is measured by the overflow current into the TIA. The result is shown in Figure 3.26. The IC demonstrates a ∼ 80% power enhancement ratio (or efficiency compared to an ideal full wave rectifier) at continuous wave case (duty cycle = 1), far below the theoretical limit of 200% that a perfect switch-only rectifier can achieve. The majority of this low efficiency is likely from the modifications implemented to have it compatible with pulsed ultrasound, but a part of it is also likely from board level parasitic (parallel capacitance to ground at the input). When the duty cycle becomes lower, the efficiency of the rectifier stabilizes around 62%, but drop to 43% when only 10 ppm duty cycle pulses are available.

The high power delivery capability is also tested electrically, with a 5 MΩ resistor in parallel with a 100 pF capacitor serving as the load. After the instruction “HP MODE” was sent, a 3 ms
high power confirmation delay is observed from the IC (Figure 3.27), and a mean power of 280 nW is measured from the high power delivery output when 0.5% duty cycle pulsed sine waves are used to emulate the high power mode ultrasound input.

3.4.3 Ultrasound Based Characterization

The Verasonics Vantage 256 system connected to an L12-3V linear array probe is programmed to deliver the designed ultrasound pulses (see Subsection 3.2.3), to test the functionality of the designed system during a real ultrasound sonography session. The waveforms generated from the linear array transducer are captured using a hydrophone (Onda Corporation) placed at the far field of the imaging transducer, and are plotted in Figure 3.28.

Here, a three cycle long pulse is used to transmit a downlink “0”, while a five cycle long pulse is used to transmit a downlink “1”. To keep the relative power level stable at the implant side, a four cycle long pulse is used when the imaging system is collecting the uplink data. The center
frequency of the ultrasound wave is, again, at 4 MHz.

With a 400 kPa ultrasound pulse hitting on the implant, the IC is powered up successfully. The captured time domain waveform for \( V_{CC} \) and \( V_{DD} \) are plotted in Figure 3.29. A power up delay of 4.7 s is measured, from the starting point of the imaging process, to the time \( V_{DD} \) reaches a desired level. This long delay is not reproducible in simulation even with testing environment replicated to the simulation testbench with author’s best effort, and is believed to be partially related to deep-subthreshold model accuracy issues. This \( \sim 5 \) s power up delay hampers the user experience when searching for such implants in an ultrasonography session.

The recorded clock and bidirectional data link during the modified ultrasound B-mode imaging process are plotted in Figure 3.30. In the captured frame, the clock is perfectly synchronized to the frame rate of 50 fps. The downlink data is also correctly recovered, and the instruction “QUERY ID” is detected. The IC then sends back the response, with its own ID of 0xFF. The missing clock edges can also be noted when the uplink data is low (or a “0”), cross verifying the fact that no clock recovery happens when the piezo’s two terminals are shorted, and justifies the complicated design choices adopted in Subsection 3.3.5.

A representative raw acoustic waveform captured on the imaging transducer (or recorded echo) is shown in Figure 3.31. After the initial pulse sent at time 0 \( \mu \)s, a reflected echo is collected after \( \sim 20 \) \( \mu \)s, indicating the implant is placed about 15 mm apart from the source transducer. Between a reflected “1” (red) and a reflected “0” (blue, behind the red curve), a pressure difference of 0.64 kPa is detected, reflecting a 19.5% change in the intensity of the echo. This becomes more obvious,
if the pressure of the echo received at a specific location is plotted across acquired frames (Figure 3.32).

From the frame-to-frame plot, a mean 700 Pa signal amplitude is detected at where the signal level is the highest, and a frame-to-frame baseline pressure variation of around 20 Pa root mean square (rms) is treated as a background noise. Based on this, the normalized energy per bit $E_b/N_0$ can be calculated, and for the specific frame displayed in Figure 3.32, this value is 24.9 dB. From communication theory, the normalized energy per bit can be directly related to bit error ratio, a measurement of the reliability of the data link. The high $E_b/N_0$ achieved in this work yields to a negligible bit error ratio smaller than $10^{-100}$, such that this data link can be treated as a digital link at the location where the implant is found.

The usage of the normalized energy per bit brings another advantage to this data transmission scheme. It is found that the spatial distribution of $E_b/N_0$ is highly correlated with the actual location of the implant from B-mode image. One example of this spatial distribution, compared to the actual

Figure 3.28: Measured ultrasound pulses from the imaging transducer, captured by a hydrophone.
Figure 3.29: Time domain waveforms for the unregulated $V_{CC}$ and regulated $V_{DD}$.

B-mode image is shown in Figure 3.33. Note that the passive layer of acoustic mismatch completely disappears in the data domain, as expected from a passive substance. The horizontal spread of the detectable signal is much higher than the raw B-mode image, likely due to the existence of side lobes in the imaging process, and a lack of well-developed algorithms like delay-and-sum [65, 74] used in B-Mode image reconstruction for SNR and accuracy boost. The 3-dB roll-off range of the normalized energy per bit region can be used as a localization method (roughly labeled with a yellow box in Figure 3.33) for on-the-fly device tracking, an idea will be revisited in Section 3.5. This finding may look obvious for now, yet its full potential will be revealed in Subsection 3.4.4.

Finally the maximum implant depth is determined by submerging the device in castor oil, and record the minimum required pressure in the ultrasound pulse at the source of the imaging transducer to establish the power delivery and bi-directional data link. Castor oil is a medium that has a similar acoustic property to that of the fat layer, one of the worst case scenarios for ultrasound energy attenuation in in vivo ultrasound applications. A higher minimum source power, as well as a degraded normalized energy per bit is recorded, as the device moves further apart from the source transducer. The B-Mode image of the implant at different depth is plotted in Figure 3.34 (A), and the recorded minimum source pressure, as well as the maximum $E_b/N_0$ achieved at different depth.
Figure 3.30: Waveforms for clock and downlink data recovered on the chip, and the uplink data from the chip under imaging mode ultrasound.

Figure 3.31: Recorded ultrasound waveforms at the transducer, when the uplink data is an “1” (red) or an “0” (blue).

are shown in Figure 3.34 (B).

In ultrasound imaging sessions, it is more likely to hit the maximum pressure limitation before the power limitation, as the temporal distribution of the pulse energy is extremely sparse. With too high a peak negative pressure (PNP), local pockets of vacuum from this pressure wave can cause mechanical damage in the tissue. This is regulated through mechanical index (MI), defined as PNP (in MPa) / $\sqrt{F_C}$ (in MHz) ($F_C$ is the center frequency of the ultrasound pulse), under a maximum allowed limit of 1.9 MPa/(MHz)$^{0.5}$. From the theoretical attenuation property of the caster oil, an implantation depth limit of 71 mm is predicted. However, achieving such a limit requires perfect angular alignment, a topic the author didn’t have time to thoroughly investigate within the limited
Figure 3.32: Pressure at a specific location in the reconstructed image, where an implant is found, plotted against the number frames during acquisition.

Figure 3.33: Example of an acquired B-mode image of a device, and its $E_b/N_0$’s spatial distribution.

amount of time. It can be also noted that the normalized bit energy rolls off fast as implant depth goes higher, yet the understanding to this problem is still very limited.

Though only 100 pW power delivery is specified as necessary in the design phase, by tuning the output pressure from the transducer, up to 1 nW excess power can be easily harvested under native imaging mode. This value is calculated by matching the voltage at the rectifier output $V_{CC}$ to the dc I-V curve from the electrical measurements. This does not only opens up possibilities for a much more complicated logic, but also give imaging mode sensing a chance, which will lead to truly distributed, parallel, and trackable bio-sensing applications.
Figure 3.34: Minimum power required from the transducer, as well as the $E_b/N_0$ as a function of implant depth.

3.4.4 Multi-Implant Operation under B-Mode Sonography

To show the capability of parallel operation of the implant, an experiment is designed where two devices are sandwiched in layers of chicken breast, submerged in castor oil, while one ultrasound B-mode imaging transducer is used to interact with both of them. The experiment setup as well as its results are shown in Figure 3.35.

In this experiment, an “QUERY ID” instruction is first broadcast, and both of the two devices respond with their IDs respectively, with one of them having its ID set to 0xFE (see Subsection 3.4.1) ahead of time. Second, the ID is used as a part of the simplest ID-specific instruction “HELLO”, and two devices replied to the instructions independently without noticeable interference. What is even more impressive, is that the locations of the two devices are hardly noticeable in the B-mode image, yet their data signature is well localized. This is the first demonstration known to the author in which individually trackable and miniaturized ultrasound based implants can operate in parallel.

3.5 Summary

In this section, to author’s best knowledge, the world’s first ultrasonography compatible power and data telemetry system is presented, opening up a new field for miniaturized, battery-less,
distributed real-time bio-sensing in vivo. However, since it is an application scenario that no previous work has ever explored, it is hard to make a comparison against prior art. Yet it is still possible to deliver a more quantitative understanding on how well this newly proposed approach works.

The proposed sensor system, on one hand, can be viewed as a novel way to track deep tissue implants. Conventional tracking strategy usually requires the implant to be battery powered, for instance the work presented in [75]. In this way, a strong signal can be delivered to the outside of the body, where multiple receivers at different locations can measure the received signal intensity and back calculate the absolute coordinate of the implant. This method is commonly referred to as received signal strength indicator (RSSI). The proposed IC, when used as an implementation of an ultrasound based tracking mechanism, offers a much higher tracking accuracy within a reasonable range thanks to the shorter wavelength compared to electromagnetic waves in common frequency ranges. Yet non-interfering with the sonography session, the location of the sensor can be bio-geographically mapped against the skin and other surrounding organs. The comparison coming from this viewpoint is summarized in Table 3.3.

<table>
<thead>
<tr>
<th></th>
<th>This Work</th>
<th>ISSCC 2018 [75]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>180 nm</td>
<td>65 nm</td>
</tr>
<tr>
<td>Battery Required</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Tracking Mechanism</td>
<td>Ultrasound Imaging</td>
<td>RSSI</td>
</tr>
<tr>
<td>Number of transducers</td>
<td>1 (commercial array)</td>
<td>8 (custom)</td>
</tr>
<tr>
<td>Implant Depth</td>
<td>71 mm</td>
<td>&gt; 300 mm</td>
</tr>
<tr>
<td>Implant Size</td>
<td>11 mm³</td>
<td>3620 mm³</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison between the proposed tracking mechanism and RSSI.

Another way to see the potential of this newly proposed implant, is by comparing it to other ultrasound powered implants. From Table 3.4, this work is the first that implements a tracking mechanism for ultrasound based sensor platforms, while still achieves a comparable implant depth and implant size without the usage of near-continuous wave ultrasound.
There is still one last question that remains. Now that the first prototype of the ultrasonography compatible implant platform shows the capability to deliver nano-watts level power under imaging mode ultrasound, is it possible to expand it for real bio-sensing applications, without resorting to the “high power mode”? If this can be implemented, locating, tracking, and sensing all can be done under the imaging mode, which will lead to truly distributed deep tissue bio-sensing applications. This is addressed in the next chapter.
Figure 3.35: Experiment setup to demonstrate parallel operation of two devices under the FoV of one imaging transducer.
Chapter 4: A Sub-nW Integrated pH Sensor

In situations where an implantable sensor platform is only able to supply a low amount of power, one question is whether it is possible to implement some kind of sensing at this level of power budget. To demonstrate this possibility, here an electrochemical pH sensor is designed that aims at sub-nW operation while delivering highly accurate sensing results. Furthermore, the robustness of the designed pH sensor can be enhanced using the extra configurability provided by the static registers in the ultrasound power and data telemetry platform. To verify the performance of the designed pH sensor, two types of the interface are designed, with one of them making it a standalone testable IC, and the other for a potential integrated operation with the ultrasound sonography compatible sensor platform presented in Chapter 3.

The verification results of the standalone sub-nW pH sensor is first demonstrated in [76].

4.1 Background

Aside of a mere demonstration of our ultrasound frontend circuits sensing potential, pH sensing itself has on its own a wide range of applications. Over the past, pH sensing has played an important role in places including but not limited to ecosystem maintenance, agriculture activity, and pharmaceutical production [77]. Recently, the rise in the concept of internet of things (IoT) has motivated researchers looking into distributed pH sensing [78], where power budgets as well as product size in such devices are usually much more constrained than traditional bench-top or handheld devices.

Another field where pH sensing is important is its usage in biomedical applications. Patients with gastroesophageal reflux disease (GERD) shows a significantly higher rate of acidic reflux in the esophagus [79], which can be detected by medical professionals using commercially available pH
sensors that are designed to be inserted down through the nostril into the esophagus [80]. However, not all pH changes within our human body are as well studied, and have well established solutions to monitor \textit{in vivo}. For instance, pH changes have been observed in wounds, and are known to have an effect on the prevention of a bacterial infection [81]. However, there is so far no known solution to offer detailed spatial pH information within the wound site, for a monitored and controlled wound healing process. The development of a standalone, miniaturized, and reliable pH sensor would significantly make the tool chest available more powerful to understand such biological processes.

4.2 System Design

To convert the concept of a low power, implantable pH sensor into a real-world piece of circuit, it is important to first translate this vague description into a set of specification, and quantify them as much as possible. Following this spirit, first, the design specifications will be introduced in this section, and a rough block level implementation of the circuit is presented. After that, the design of each individual circuit will be explained in detail, which largely rely on the design flow introduced in Chapter 2.

4.2.1 Top Level Design

Conceptually and qualitative speaking, a pH sensor can be roughly broken into four parts, as is shown in Figure 4.1. A sensor is required to convert the chemical signal into an electrical one, which is usually analog. An analog front end (AFE) is followed up for signal conditioning purposes, which is further followed by an analog-to-digital converter (ADC). Finally, the output of the ADC is processed in the digital domain to interface with other building blocks that may be placed outside of the chip. One or more blocks followed by the sensor can be omitted depending on the specific application scenario, for instance, the entire chain after AFE can be removed if the sensor’s output is directly digital, and the ADC can be removed in rare cases when analog output is preferred.

From a “black-box” level point of view, there are certain performance metrics that needs to be optimized for this pH sensor. The first is that it should deliver a pH readout with high accuracy.
This is ultimately dominated by the signal-to-noise ratio (SNR) of the front end sensor, but it may be degraded when it passes through AFE and ADC for their added noise and distortion. The second is that the pH sensor circuit should be able to deliver such an output at a reasonable speed. The speed of change for pH in common biological environments are generally low, as the settling time to a new equilibrium from a local hydronium ion concentration change usually requires lengthy period for ions to diffuse. Sensing mechanism itself may also add extra time constants, and the speed of choice should accommodate these factors. The third concern is on the power consumption. In this case, a sub-nW total power consumption is targeted, partially to demonstrate the sensing capability under sub-nW regime for distributed biosensor platform proposed back in Chapter 3. The final concern is the area. A smaller area will make room to reduce the size of the overall implant. This translates to a reduced mechanical damage during its long term operation after implantation.

Four metrics have been defined from the previous discussion. Among them, one can be fixed, which is the speed of operation. The sampling speed target for this design is set to 1 sample/s, a common choice in commercial pH sensors. The major advantage from this sampling speed is that this keeps the user less paranoid with suspicions that the device may not work. This is certainly a concern for the testing personnel, such that a sampling speed of 1 sample per second is chosen to maximize its tolerance to the impatience of the user (happens to be the author of this thesis). Other than the speed, the other three metrics should all be optimized. The requirement for all four metrics are summarized in Table 4.1.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Accuracy</th>
<th>Speed</th>
<th>Power Consumption</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimize for</td>
<td>High</td>
<td>1 Sample/s</td>
<td>Low</td>
<td>Small</td>
</tr>
</tbody>
</table>

Table 4.1: Desired performance of the designed pH sensor

It is also important to investigate what extra constraints a potential integration with the ultrasound sensor will impose on top of the general consideration of a pH sensor. The first constraint is the maximum clock frequency. If the ultrasound interface is designed to be synchronized with a 50 Hz clock, then the pH sensor should also operate at a clock frequency that is an integer division of 50 Hz. From this, a nominal clock frequency of 25 Hz is chosen. The second constraint is the on chip power supply. Aside from a limited power available, the on chip supply voltage is also not so well defined, and is only known to be larger than 1.2 V. Extra power regulation is required for sensitive analog circuits. Fully differential topology is also preferred as it has an intrinsic advantage in PSRR. There are more subtle constraints to be added here, however, they are easier justified once the design of certain blocks becomes clear. The issue of interface compatibility to the ultrasound sensor platform will be revisited once the design of the core circuit is thoroughly introduced, in Subsection 4.2.7.

### 4.2.2 ISFET and REFET

The most important building block for the integrated sensing system is the sensor itself. Here we chose to use the ion-sensitive field effect transistor (ISFET) as the sensor that converts the H$_3$O$^+$ ion concentration into an electrical, analog signal, for its native compatibility with commercial CMOS fabrication process that leads to low cost implementations.

ISFET was first introduced in 1968 under a different name, and its performance was not fully verified until 1972 [82, 83]. The way ISFET senses the pH changes in the environment is from a voltage difference that is generated by the electrochemical equilibrium established between the gate oxide layer and the solutions that have different concentration of H$_3$O$^+$ ions (or pH).

With a slight modification to the topology, the first integration of an ISFET and conventional CMOS circuits is published in [84], to the author’s best knowledge. With rapid advancement
of the CMOS technology, ISFET becomes widely adopted for integrated ion sensing applications, covering applications like traditional pH sensing [85, 86], gas sensing with micro-electromechanical system (MEMS) integration [87, 87], genome sequencing with modified surface properties [88, 89], and many more. It is also chosen as the chemical sensing frontend in this design to fulfill a low power pH sensing IC, as minimal and mostly predictable difference appears between the design and the actual, physical device, thanks to the almost 50 years of effort on understanding the behavior of ISFETs, and a maximum translation of skills, as a circuit designer, to the optimization of the sensor’s performance.

The principle of an integrated-circuit-compatible ISFET’s operation can be understood as the following. For the passivation layer materials that are commonly used in the IC technology, silicon oxide (SiO$_2$), silicon nitride (Si$_3$N$_4$), and silicon oxy-nitride (SiO$_X$N$_Y$), all have well defined association-dissociation reactions with hydronium ions (H$_3$O$^+$) in an aqueous solution, defined as [90, 91, 92]:

\[
\begin{align*}
\text{SiOH}^- + \text{H}_3\text{O}^+ & \rightleftharpoons \text{SiOH}_2 + \text{H}_2\text{O} \\
\text{SiOH}_2 + \text{H}_3\text{O}^+ & \rightleftharpoons \text{SiOH}_3^+ + \text{H}_2\text{O}
\end{align*}
\]

for surface silanol sites (SiOH$^-$), and:

\[
\begin{align*}
\text{SiNH}_2 + \text{H}_3\text{O}^+ & \rightleftharpoons \text{SiNH}_3^+ + \text{H}_2\text{O}
\end{align*}
\]

for surface amine sites (SiNH$_2$). The equilibrium of these reactions moves as the concentration of the hydronium ions changes. When the concentration of H$_3$O$^+$ is high (a low pH value), an accumulation of surface charge will happen as the equilibrium moves to the right hand side to minimize the Gibbs free energy of the overall system. A higher surface charge density will then cause a higher surface potential, as illustrated in Figure 4.2. This surface potential $\Phi_S$ is to the first
order, governed by the Nernstian equation [83, 93]:

\[
\Phi_S(pH) = \ln(10) \phi_T \alpha \left( pH_{pzc} - pH \right)
\] (4.3)

where the pH\(_{pzc}\) stands for the pH at the point-of-zero-charge, and the attenuation factor \(\alpha\) is expressed as:

\[
\alpha = \frac{q \beta_S}{\ln(10) \phi_T C_S + 1}
\] (4.4)

Figure 4.2: An illustration of the sensitivity of the surface charge to the change in pH. SiO\(_2\) passivation is used as an example.

where \(\beta_S\) is the surface buffer capacity, i.e., the ability to absorb or release protons, and \(C_S\) is the differential (small-signal) double-layer capacitance. Within a reasonable range of pH before the surface buffer capacity is close to saturation, it can be seen from Equation 4.3 that the surface potential is roughly in a linear relationship with pH of the solution.

Now if a MOS transistor is placed directly underneath the exposed passivation, with its gate electrically extended to as close as possible to the sensing surface, this change in surface potential can be translated into a change in the gate voltage of the MOSFET, through an extra capacitor divider. However, to sense this voltage change, an electrical connection to the solution is required.
This can be implemented by attaching a reference electrode to the solution, which maintains a constant voltage offset between its electrical terminal and chemical terminal through a material-defined electrochemical reaction. However, a “true” reference is usually bulky and expensive, for its incorporation of a porous glass based diffusion barrier. This diffusion barrier is in place to define the equilibrium point of the reference electrode reaction, as Nernstian equation predicts that all electrochemical equilibria are sensitive to the concentration of the species involved.

Having a “true” reference electrode increases both the size of the implant, and the cost of the overall system. However, a simple solution to this problem exists. REFET, which is a local, matched duplicate of the ISFET with its sensing site blocked, can be used to monitor the drift from the externally placed pseudo reference electrode. The differential signal between the ISFET and the REFET pair is now defined as the electrical output of the sensor. A conceptual drawing of this
structure is shown in Figure 4.3.

Now that the requirement for the reference electrode is relaxed, a quasi-reference electrode (QRE) can be used, whose potential changes weakly it is in contact with different solutions. In circuit terminology, the ISFET-REFET structure translates the effect of having an non-ideal reference electrode into a common mode signal, while the chemical information is encoded in the differential mode signal.

It is worth noting that the deposition of additional coating layer, as well as having a floating gate in a MOSFET, may introduce a random amount of floating charges that effectively shifts the gate potential. To accommodate this risk, in this design, both NMOS-based and PMOS-based ISFET/REFET pair are implemented as different chips, while keeping the signal processing circuitry identical. The QRE can be connected to either the power or the ground rail for a better common mode voltage, giving four possible common mode voltage and sensor configuration combination. The best setup that has the most favorable performance can then be selected. In the following subsections, schematics with NMOS ISFETs only will be shown, for simplicity without a loss of generality.

With the sensing principle well understood, it still remains a question that how this signal can be reliably amplified and digitized. This is what the rest of this section covers, starting from the circuitry that directly interacts with the ISFET/REFET pair, the ISFET frontend.

4.2.3 ISFET Frontend

The ISFET/REFET pair can be treated as a capacitive coupled sensor between the chemical domain and the electrical domain, since the ion concentration in the chemical system is modulating the gate charge. However, what makes it different to a traditional capacitively coupled sensor, is that on the other side of the gate, three terminals exist. It is possible to utilize this intrinsic topological difference for novel compensation schemes that gives a boost in its performance, however, it is more important to make sure that this flexibility does not introduce extra undesired performance degradation.
In the spirit to make it a better capacitive sensor frontend, the controlled-voltage-controlled-current (CVCC) circuit was proposed [94, 94, 95]. The idea may be published under different names, but the idea remains largely the same. Since the drain current of the ISFET is uniquely defined by the voltage on its all four ports, which can be expressed as:

\[ I_D = I_D(V_{GS}, V_{DS}, V_{BS}), \] (4.5)

if the body-to-source voltage \( V_{BS} \) is fixed by physically connecting the body to the source, and somehow the drain-to-source voltage \( V_{DS} \) is also fixed, then the drain current \( I_D \) becomes uniquely defined by \( V_{GS} \), and vice versa: a constant \( I_D \) in this case gives a constant \( V_{GS} \). This idea is illustrated in Figure 4.4 (A). One CVCC implementation is shown in Figure 4.4 (B), where two amplifiers are used. With the drain current through the ISFET fixed by the bottom current source, the source voltage is buffered and copied to a separate branch. In this branch, another fixed current passes through a fixed resistor to generate a fixed voltage, which is then added on top of the buffered source voltage. This new voltage is buffered back to the drain node, thus a fixed drain-to-source voltage is established. In this case, this fixed drain-to-source voltage is \( I_{DS}R_{DS} \). Variance to this circuit exists when scaling down to low power domain. One notable example is [96], in which ADCs and DACs are used to replace the buffer amplifiers, such that the resistor \( R_{DS} \) can be removed and replaced by a digital configuration circuit for any desired drain-to-source voltage.

However, no matter what is used to implement the CVCC principle, it is obvious that the majority of the circuit blocks are in place to generate a relatively fixed \( V_{DS} \). Although these implementations work well in principle, they add extra area, power, and noise directly to the implementation of the sensing frontend. On the other hand, when biased in deep-subthreshold, the drain current has an intrinsic low dependency on the drain-to-source voltage under subthreshold saturation. Thus here, a pseudo-differential source follower pair (Figure 4.5) is proposed to translate the gate charge, or gate voltage difference to the source node for sensing with a moderate accuracy. However, to build a full system, it is important to know quantitatively how much distortion it would introduce from
not fixing \( V_{DS} \). Or more specifically, in the pseudo-differential source follower pair, if there is a \( \Delta V_G \) potential difference in the gate voltage between the ISFET and REFET, how much nonlinearity will this structure add to the output differential mode signal \( \Delta V_S \)? It is possible to perform both theoretical analysis and numerical simulation to better understand this issue.

Using Equation 2.3, assuming the ISFET and the REFET have the same size, and pass the same amount of drain current, then:

\[
I_D = I_S \frac{W}{L} e^{(V_{GS,IS} - V_{TH})/\phi T} \left(1 - e^{-V_{DS,IS}/\phi T}\right) = I_S \frac{W}{L} e^{(V_{GS,RE} - V_{TH})/\phi T} \left(1 - e^{-V_{DS,RE}/\phi T}\right)
\]

\[
e^{(V_{GS,IS} - V_{GS,RE})/\phi T} = \frac{1 - e^{-V_{DS,RE}/\phi T}}{1 - e^{-V_{DS,IS}/\phi T}}
\]

To further simply this equation, it is important to take the following assumptions. First, the gate voltage of the REFET \( V_{G,RE} \) is approximated as a constant, assuming the drift on the QRE is small. Its source voltage \( V_{S,RE} \) is now also a constant, since its sensing site is blocked and its gate-to-source voltage is not sensitive to pH changes in the solution. This approximation leads to
Figure 4.5: Topology of the proposed pseudo-differential source follower pair.

several conclusions. The output common mode voltage can be denoted as $V_{DS0} = V_{DS,RE}$ for the drain-to-source voltage of the REFET at an calibrated pH value, where $\Delta V_G = 0V$. The value $V_{DS0}$ here is approximately a constant, since the drain voltage of both the ISFET and the REFET is $V_{DD}$, and the source voltage of the REFET is considered as a constant. Now assume $V_{DS0} >> \phi_T$, and define $\kappa = e^{-V_{DS0}/\phi_T} << 1$, another constant that helps simplifying the analysis. With the above notations and approximations, Equation 4.6 can be rewritten as:

$$\Delta V_G = \Delta V_S - n\phi_T \ln \left( \frac{1 - \kappa e^{\Delta V_S/\phi_T}}{1 - \kappa} \right)$$

(4.7)

Up to this point, the equation contains only two variables: $\Delta V_G$, which is the input, and $\Delta V_S$, which is the output, with everything else being constants (approximately, not a function of terminal voltages). Distortion becomes obvious if an analytical expression of $\Delta V_S(\Delta V_G)$ exists. However, this requires the help from the Lambert W function, and the resulting expression offers limited intuition to circuit design choices. The inverse relationship, $\Delta V_G(\Delta V_S)$, is investigated here instead. And Taylor’s expansion is applied on Equation 4.7 to generate intuitive understandings on
the harmonic distortions based on the approximations taken so far:

\[
\Delta V_G(\Delta V_S) \approx \Delta V_S - n\phi_T \left(0 - \frac{\kappa}{\phi_T} \Delta V_S - \frac{\kappa^2}{\phi_T^2 (1 - \kappa)} \Delta V_S^2 + O(\Delta V_S^3)\right)
\]

\[
= (1 + n\kappa)\Delta V_S - \frac{\kappa^2}{1 - \kappa \phi_T} n \Delta V_S^2 + O(\Delta V_S^2)
\]

(4.8)

Ignoring third order and higher order distortions, this is roughly:

\[
\Delta V_G(\Delta V_S) \approx (1 + n\kappa)\Delta V_S - \frac{n\kappa^2}{\phi_T} \Delta V_S^2
\]

(4.9)

From Equation 4.9, an attenuation factor of \(1 + n\kappa\) is found from the input \(\Delta V_G\) to the output \(\Delta V_S\). Also, the smaller \(\kappa = e^{-V_{DS0}/\phi_T}\) is, the lower the attenuation, as well as the second order distortion (in fact, this is also true for higher order terms). This means a higher \(V_{DS0}\) is generally desired for a higher signal-to-distortion ratio.

While the analysis provided above delivers limited intuition, numerical simulation provides no more. Yet it is a better way, and the most accurate way known, to quantify the actual amount of distortion in \(\Delta V_S(\Delta V_G)\). To perform the numerical simulation, the gate voltage of the REFET is adjusted such that the source voltage is exactly one-half of the \(V_{DD}\), while the ISFET’s gate is swept in \(\pm (1/2V_{DD} - 100\text{ mV})\) range. This range is chosen to capture the distortion effect all the way before one of the transistors exit subthreshold saturation. A simulation result with \(V_{DD} = 800\text{ mV}\) is shown in Figure 4.6.

With the best linear fit subtracted from the \(\Delta V_S(\Delta V_G)\) simulation result, the sum of all higher order distortion terms can be then quantified. It turns out that the transistor model predicts a maximum absolute distortion \(|\epsilon|\) less than 0.03% of the total simulated range. Although it might be a good idea to be skeptical about the model accuracy in this case, from what was discussed in Section 2.1 and Section 2.2, it seems more likely that the numerical model, if inaccurate, is leading to a higher dependency of \(I_D\) to \(V_{DS}\), and predicts a higher distortion than what would have been predicted by the simplified analytical model based on Equation 2.3.

A bounded \(\pm 0.03\%\) distortion means that the signal coming out from the proposed pseudo
differential source follower pair is good enough to deliver a 10-bit resolution data within the working range, while leading to a massive power and area reduction compared to CVCC-based solutions.

And thus, a 10-bit final output resolution is targeted for the end-to-end accuracy of the designed pH sensor, subject only to degradation from the quantization process that is intrinsic to all analog-to-digital conversions.

4.2.4 ADC

With a targeted 10-bit resolution, the ADC is designed in a minimum-effort style. A 10-bit successive approximation (SAR) topology is chosen, for its relative simple topology, and widely reported great performance in nW-power regime [97, 98, 99, 100, 101]. The schematic of the SAR ADC is shown in Figure 4.7. The main challenge to translate nW-regime ADCs into sub-nW ones is that leakage energy now becomes by no means negligible, thus degrading the Walden figure-of-merit (FoM) [102], a measure of energy efficiency per bit information.

In this design, a fully binary-weighted capacitor array is used, implemented using MiM capacitors available in the technology (with a density of 2 fF/µm²). A minimum sized MiM capacitor is used for the unit capacitor, with a mean capacitance of 35.6 fF. To minimize routing induced parasitic mismatch, Metal 4 is used as a shielding layer that is electrically connected to the top plate, and the bottom plate is used for sampling and holding (Figure 4.8). The addition of the shielding layer adds an extra ∼ 3.1 fF capacitance predicted by post layout extraction. Switching capacitance
dominates the dynamic power consumption in most cases for a SAR ADC, thus a smaller sized unit capacitor is usually preferred. Techniques like split capacitors [103] has gained popularity in low power designs like [99, 100] for an exponential decrease in the switching capacitance, however, its impact in linearity [104] is non-negligible, since a fraction-sized attenuation capacitor needs to be made in a matched, accurate way. This introduces a reduction in the yield for a given desired linearity, feasible for standalone ADCs if an extensive amount of testing is used to filter good chips out, but not quite so for fully integrated system without separate testability built in.

To reduce the switching energy used by the fully binary capacitor array, a monolithic switching scheme is adopted [105], such that the capacitors are only charged once for every sample, and only discharges when successive approximation comparison happens. And the reduction of the
switching energy comes with an minimum addition to the complexity of the digital SAR control logic. However, one significant feature this introduces is a varying common mode voltage present at the input of the differential comparator. For a monolithic down-switching process, the common mode voltage at the input of the comparator can go from the sampled common mode $V_{CM}$ all the way to the ground rail $V_{SS}$. It is thus important to minimize the variation in the input-referred mismatch of the comparator across common mode voltages between $V_{SS}$ and $V_{CM}$, and keep the input-referred noise level under control within this range.

![Figure 4.9: Schematic of the comparator used in the ADC.](image)

To achieve a minimum degradation in the ADC’s performance, the comparator is designed in a way similar to what is originally presented in [105] (schematic shown in Figure 4.9). With a PMOS input transistor pair biased under controlled drain current, the $V_{GS}$ and the $g_m$ of the input transistors are fixed, leading to a roughly constant mismatch and noise behavior across different input common mode voltages. The input-referred comparator noise is dominated by the parasitic capacitance at node $V_{X+}$ and $V_{X-}$ [106, 107] in a constant scaled $A kT/C_X$ form. To get a desired noise behavior, two matched 12 fF capacitors are attached to the node $V_{X+}$ and $V_{X-}$, implemented using a MoM-like layout from Metal 2 to Metal 4 (Figure 4.10). The choice of the tail current then determines the speed of this comparator. To operate with 25 Hz clock (see Subsection 4.2.1), a 60
pA tail current is chosen, which incorporates a margin reserved for anticipated model inaccuracy.

![Diagram of the comparator used in the ADC](image)

Figure 4.10: Layout of the comparator used in the ADC.

4.2.5 Switched-Capacitor Amplifier

A switched-capacitor amplifier is added in between the ISFET frontend (Subsection 4.2.3) and the ADC (Subsection 4.2.4), since the ISFET frontend circuit’s high output impedance cannot drive the sampling “bottom plate” capacitor of the ADC at the desired speed of one sample per second. On top of that, there are two more benefits from adding a stage of switched-capacitor amplifier. The first benefit is that it implements a common mode voltage translation. The common mode
voltage at the output of the ISFET frontend is signal-dependent, however, with a proper common
tode feedback circuit, the common mode voltage at the output of the switched-capacitor amplifier
can be fixed to a certain value, and the mid-rail voltage $1/2V_{DDA}$ is chosen in this design. This is
important to ensure the comparator in the ADC meets the speed requirement. The second benefit
is that a differential mode voltage gain can be implemented to further utilize the input voltage range
of the ADC. A gain of 2 is chosen here to guarantee that even with the highest input signal swing
possible (which is a near-Nernstian response from the ISFET), the pH sensor can still cover a range
of 7 pH variation.

![Schematic of the switched capacitor amplifier.](image)

Figure 4.11: Schematic of the switched capacitor amplifier.

The schematic of the switched-capacitor amplifier is shown in Figure 4.11, where the transistor
level design of the fully differential amplifier is shown in Figure 4.12 (A). The sizing in the
differential path largely follows what was presented in Subsection 2.2. The common mode feedback
is implemented through a sample-and-hold fashion (Figure 4.12 (B)). During the sampling phase
of the switched capacitor amplifier, all four differential input and output terminals are shorted
together, and the common mode feedback loop adjust their voltage to the input $V_{CM} = 1/2V_{DDA}$.
During the hold phase, all shorting switches open, and two matched capacitors are used to track the
output common mode during this phase. Finally a current mode miller feedback is implemented
by a capacitor $C_C$ to the source of the differential input pair to create a left half plane zero in the common mode loop gain, enhancing the stability of the common mode feedback loop. The value of this capacitor is experimentally determined to suppress the common mode ringing to an acceptable value, with extra added margin for potential stability degradation from model inaccuracies.

![Figure 4.12: Schematic of the fully differential amplifier.](image)

Chopper stabilization [29] is not adopted in the amplifier design, as the highest frequency clock available is only 25 Hz. Not only does chopping at this frequency provides a limited noise reduction, a filter that preserves the 1 Hz signal bandwidth yet provides enough suppression for the 25 Hz chopping artifacts is also hard to implement.

The input capacitor and the feedback capacitor forms a ratio of 2 to implement the gain. Their absolute value are determined by the speed requirement and the output impedance of the ISFET frontend stage.
4.2.6 Periphery Circuit Blocks

To support the core analog signal chain, several periphery circuit blocks are required. The first is the circuit block that generates the common mode voltage $V_{CM} = 1/2V_{DDA}$. The schematic of the common mode generator is shown in Figure 4.13. It can be viewed as a buffered voltage divider, where the buffer’s driving capability (or its output impedance and slew rate) is characterized and sized such that it creates a negligible common mode error at all times during the switched-capacitor operation.

Another important block is the current source. The current source is responsible to bias all analog circuits, including the comparator in the ADC, since all of their performances are defined by their bias currents. In this work, a simpler approach is taken instead of building an on chip temperature-compensated current reference. Shown in Figure 4.14, the current source has a matched array of gate-source connected PMOS’s. Through an externally supplied “current code”, the output current can be tuned digitally, not only to achieve the desired performance, but also provide adjustment for model inaccuracies. When integrated with the ultrasound sensor platform, this current code can be easily sent through the defined protocol to the configuration registers. Thus, the current code approach won’t lead to too much of a complication in the targeted
application scenario. Another advantage of this solution is that it provides a simple adjustment interface when testing under different environments. In the targeted application scenario, which is pH sensing \textit{in vivo}, the temperature of the working environment is a stable value around 37°C. While in lab-based bench-top testing, the surrounding temperature is usually between 20°C and 22°C. Since the subthreshold current is sensitive to temperature changes, this simple solution allows easy adjustment for both environments, without much of a power overhead usually present in temperature-compensated current references.

![Diagaram showing a digitally configured current source](image)

Figure 4.14: The digitally configured current source.

A current mirror is used to distribute the current to different analog blocks.

For the integrated version, an additional voltage regulator is required to generate a quiet analog supply $V_{DDA}$ from the noisy, unregulated source supply voltage. This is done by using the same regulator design with another local copy of the voltage reference used in Subsection 3.3.2. An additional 2-bit configuration is embedded to select which node in the resistor ladder (see Figure 3.12) is fed back to the input of the OTA, effectively offering a 2-bit adjustable output voltage. It is important to notice that a change in the analog supply $V_{DDA}$ does not change the power consumption of the analog circuits in this pH sensor to the first order, since almost all circuits are current-mode biased and this current is ultimately supplied through a linear regulator. The adjustment on the
analog supply $V_{DDA}$ only provides an on-the-fly tuning capability between a higher accuracy of the sensor (when a lower supply voltage is chosen) and a higher sensing range (when a higher supply voltage is chosen). A change in current code will scale the power consumption, but will have an impact on the speed of operation.

One subtle issue exists when powering the pH sensor up in an integrated scenario: the power up behavior introduces a charge sharing between the storage capacitor in the ultrasound sensor frontend and the analog supply decoupling capacitors. The latter is usually large to preserve the clean analog signals, yet the former cannot be increased too much for a fast startup response, allowing faster device identification for the user operating the system. Thus the charge sharing between the two is likely to cause a massive voltage droop in the ultrasound interface, and may interrupt the communication between the host and the implanted device. To minimize this voltage droop, the powering up of the analog supply domain of the integrated pH sensor is divided into a power up sequence of 3 sub-domains, with gating PMOS transistors in between. When a command is received from the ultrasound frontend that powers up the pH sensor, the 3 sub-domains will get charged up sequentially. Furthermore, programmable power-up delays are implemented using the static configuration registers, to better control the amount of voltage droop seen by the ultrasound interface.

4.2.7 Digital Interface

The digital interface is different between the standalone pH sensor circuit and the pH sensor integrated with the ultrasound interface.

For the standalone version, the goal of the digital interface is to reduce the total number of I/O pins. Here, the major functionality of the digital interface is to serialize the output data from the 10-bit ADC, and to provide a serialized input for current code configuration (see Subsection 4.2.6). The serial input is implemented using a scan chain, adding SE, SI, SO into the required pin list. The serialized data output is decorated with additional header and footer (parity check bits) for frame identification. A total 16 bit frame is sent for one sample, that takes 25 clock cycles. The
composition of each data frame is shown in Table 4.2, where “P” stands for parity check (the XOR, or the 1-bit addition, of the 10-bit ADC value).

<table>
<thead>
<tr>
<th>Bit #</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4 - 13</th>
<th>14</th>
<th>15</th>
<th>16 - 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>ADC Value [9:0] (Payload)</td>
<td>P</td>
<td>P̅</td>
<td>Idle</td>
</tr>
</tbody>
</table>

Table 4.2: Frame structure for the output serial data in the standalone implementation.

For the integrated design, data does not need to be serialized within the pH sensor, as this will be handled by the link layer in the ultrasound interface (see Section 3.3.7). Only very simple logic is required to run within the pH sensor module, and for simplicity, the supply of the digital logic is connected to $V_{DDA}$. But this design choice requires additional level shifting when communicating between the pH sensor and the digital circuit in the ultrasound frontend. To make the matter worse, the actual value of $V_{DDA}$ is tunable, and can be higher or lower than the digital supply on the ultrasound interface side. To solve this problem, all digital signals that traverses across the boundary between the pH sensor and the ultrasound interface is shifted to the highest value generated on the chip, i.e., the unregulated $V_{CC}$ (see Subsection 3.3.1). An I/O bank is designed to perform this task.

The mapping of the static registers in the ultrasound frontend for the static control signal of the pH sensor is detailed in Table 4.3. These can be accessed using store instructions on the fly (see Subsection 3.3.7), if the number of available configuration registers is expanded to 4.

<table>
<thead>
<tr>
<th>Address</th>
<th>Bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x03</td>
<td>CURRENT_CODE [7:0]</td>
</tr>
<tr>
<td>0x02</td>
<td>ASC[1:0]</td>
</tr>
<tr>
<td>0x01</td>
<td>POWER_UP_DELAY[5:0]</td>
</tr>
<tr>
<td>0x00</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 4.3: Register mapping for the pH sensor control signals.

In Table 4.3, the signal “CURRENT_CODE” is the digital control signal for the current source described in Subsection 4.2.6, and “ASC” stands for “ANALOG_SUPPLY_CONTROL”,

107
that changes the value of the regulated $V_{DDA}$. A counter is added to down sample the clock, and the value of the clock divisor is “CLOCK_DOWN_SAMPLER”. Thus when this value is 0x01, a 25 Hz clock is fed into the pH sensor if the ultrasound front end is synchronized to a 50 fps imaging session. “POWER_UP_DELAY” is a 6-bit value that controls how many clock cycles are waited between each of the 3 phases of sequential power-up. And finally, “APD” stands for “AUTOMATIC_POWER_DOWN”, a flag determines if the pH sensor will be powered down after every measurement. The default values (the values after the power-on reset) for these registers are: [0x03]: 0x20 (CURRENT_CODE = 0x20); [0x02]: 0x41 (ANALOG_SUPPLY_CONTROL = 0x01, CLOCK_DOWN_SAMPLER = 0x01); [0x01]: 0x40 (POWER_UP_DELAY = 0x10, AUTOMATIC_POWER_DOWN = 0x00); [0x00]: 0x00.

4.3 Performance Verification

The proposed IC (within chip set USTAG V2.7) is fabricated in the TSMC 180 nm 1P6M Salicide with AlCu FSG process (MSRFGII) technology. The verification process is done only with the standalone version of the pH sensor at the time of writing, though many design choices were dominated by the potential integration with the ultrasound sensor platform.

The standalone integrated pH sensor IC occupies a $1020 \mu m \times 830 \mu m \approx 0.85 \text{ mm}^2$ area. A die photo is shown in Figure 4.15, where the physical implementation of each block and the logic connection to each of the I/O pins are labeled in the photo. A significant amount of area is used to create on-chip alignment marks. Excluding the alignment marks, the core functional blocks occupy only $890 \mu m \times 520 \mu m \approx 0.46 \text{ mm}^2$. The core functional blocks include the ISFET and REFET front end, the switched-capacitor amplifier, the SAR ADC, the serial interface, the current bias, and the common mode voltage generator. The two variants of the IC (PMOS based ISFET and NMOS based ISFET) were copied once to make a stripe of four chips, that sits within an array of other ICs in the chip set USTAG V2.7. The entire IC array was used directly for post fabrication before dicing, to minimize artifacts in the fabrication process that is more pronounced when smaller chips are being processed.
4.3.1 Post Fabrication and Packaging

The goal of the post fabrication is to shield the REFET’s sensing area, thus making it insensitive to pH changes in the solution that it is in contact with. The fabrication flow is illustrated in Figure 4.16, primarily carried out by Filipe A. Cardoso, the second author of [76]. The process starts with the preparation of pocket carriers which are made to physically extending the dimension of the chip, which will help avoid non-uniformity that typically occurs during spin-coating, deposition and etching processes at the edge of the chip. To do this, squares with the exact dimensions of the die (chip set USTAG V2.7) is first patterned on 4-inch diameter, 500 µm thick silicon with 5 µm silicon dioxide wafers (Figure 4.16, A-B). Cavities are then etched by the Bosch process (Oxford PlasmaPro 100 Cobra ICP, C), which is precisely calibrated against process parameters to create a depth that matches the thickness of the die (300 µm). The wafer is then diced into 15 mm × 15 mm pieces with one cavity each in the middle (D). The chip set USTAG V2.7 die is then glued into the cavity using epoxy NOA 86H (Norland Products), under 5 N pressing force over the chip surface at
125 °C for 20 minutes (E). At the end of this process, the CMOS die is perfectly flat, and the top surface is at the same level as the pocket carrier (F).

The carrier with the CMOS die is then subject to Parylene coating process. A silane A174 is first deposited using vacuum evaporation process over the carrier to promote the adhesion to the Parylene. A layer of 1 µm Parylene is then evaporated (SCS Labcoter 2 Parylene Deposition System, G). 5 µm thick positive photoresist AZ9260 (MicroChemicals) is used to define the pattern of where the Parylene will stay, followed by a O₂/CHF₃ dry etching (Oxford PlasmaPro 100 Cobra ICP) process to remove the excess Parylene (H-I). Finally, the chip array is protected with another layer of photoresist and diced (Disco DAD3220 dicing saw), such that a pair of NMOS based pH sensor and PMOS based pH sensor is isolated as a chiplet (J-K). A Remover PG (MicroChemicals) soak is needed to remove the protective photoresist at the very end.

Figure 4.16: Illustration of the post fabrication flow.

An 16 pin package PCB is designed such that the chiplet can be attached to and wirebonded for electrical connections. The design of this PCB is detailed in Subsection 4.3.2. Full encapsulation is
usually used to protect the wirebonds, however, in this work, the part of the ISFET and the REFET, or the sensing site, needs to be exposed. Since the bondwires are lifted about 250 µm above the chip surface, post fabrication based methods that works with thin films cannot solve the packaging problem, although additional post fabrication steps may help if it was done prior to the dicing. Instead, a “doughnut” epoxy fill is performed manually here. To increase the chances of a success encapsulation, the following steps are used.

1. An epoxy line is drawn either between the bond wire and the sensing site, or on the empty area on the package near the sensing site, using the “dam epoxy” LOCTITE ECCOBOND FP4451.

2. This epoxy is cured on top of a tilted hot plate preheated to 165 °C for 1 minute. The sensing site is placed closer to the top of the slope.

3. The package and the chip are cooled down to room temperature, such that the epoxy applied next won’t be heated up immediately, which will lead to a lower viscosity of the epoxy, and a poorly defined line in the next step.

4. Repeat step 1-3 until the sensing site is surrounded by an epoxy dam that is approximately 0.5 mm high.

5. Draw an outer epoxy dam on the package.

6. Put the package and the chip on a preheated 60 °C hot plate placed on a flat surface. Fill the “doughnut” region between the two epoxy dams with “fill epoxy” LOCTITE ECCOBOND FP4450. The elevated temperature helps the flow of the “fill epoxy”, such that overfill is minimized which can cause epoxy overflows that will potentially cover the sensing area at the center during the actual curing process.

7. Bake the epoxy in a convection oven for 30 minutes under 125 °C, followed by 90 minutes under 165 °C.
Pictures of several key steps shown in Figure 4.17. A typical encapsulation will take around 3 hours of manual application of epoxy, followed by 2 hours of full curing. However, the yield of this entire process is still limited, with occasional broken wirebonds, and accidentally covered sensing sites.

Figure 4.17: Illustration of several key steps during the encapsulation process. (1) application of a line of epoxy, (2) epoxy after curing, (3) the full inner dam after curing, and (4) the full encapsulation.

4.3.2 PCB Design

The PCB system is designed to assist the pH measurement experiments as the intermediate hardware between the custom-designed IC and a commercial computer, to record the output digital value from the sensor chip in real time. Two extra factors add into the complexity of the designed PCB system. The first consideration is that the IC is sensitive to light, which means during an experiment, the IC needs to be optically isolated from the lab environment, where at least some
light is required to find which pH solution will be added. To solve this, the PCB based system is designed consisting two parts connected using flexible cables, such that one part can be put into a dark, isolated environment. The second consideration is that the post fabrication and the packaging together gives a limited yield, which sometimes can only be tested upon the addition of pH buffers. To reduce the work required when a failed chip is encountered, the package board is designed to incorporate only necessary buffers on the back side of the attached IC. And mechanical connectors are used to connect the package to another board for any signal processing that has some parasitic tolerance.

Figure 4.18 illustrates the functionality of the 3 boards in a block level fashion, plus a commercially available FPGA Opal Kelly XEM6010 module (green). All commercial ICs are powered at 3.3 V supply voltage, which is regulated from a 5 V input using Analog Devices LT3080 (a former Linear Technology product) on the orange board, except for the power regulators and the chip side of the level shifters. On the package board (red), all output digital signals from the custom IC are level-shifted using comparators (Texas Instrument LPV7215) by comparing their value against the half-rail voltage, to generate 3.3 V buffered digital outputs. On-package transimpedance amplifiers are designed using Texas Instrument OPA4322 amplifiers, and jumpers are in place to select one of the four high-value resistors into the loop for an adjustable transimpedance gain. The sub-pA input bias current prevents the OpAmps from injecting too much error into the read out of sub-nA current consumption from the chip.

Figure 4.18: Block diagram of the 3 boards designed for real time pH measurements.

The blue daughter board connected directly with the package hosts level shifters for input digital signals. An ADC (Texas Instrument ADC121S021) is used to digitize the transimpedance
amplifier’s output. All connections between the blue daughter board and the orange mother board are either digital or power signals, and can run safely on 12 position flexible flat cables (FFC). Power signals use more than one connections to reduce their path resistance as an extra precaution. But the overall power consumption from the chosen commercial ICs makes it unlikely to cause any trouble even when the power is delivered through Ohms of series resistance.

Figure 4.19 shows each of the individual PCBs that are designed for pH measurements. And Figure 4.20 shows their final form when everything is connected together.

![Figure 4.19: Soldered PCB’s for the pH measurement system.](image)

![Figure 4.20: Soldered PCB’s for the pH measurement system when the boards are connected as designed.](image)

4.3.3 FPGA and Software Design

The signal processing on the FPGA and the software are co-designed to deliver a debug-friendly software-controlled hardware system. The idea to hardcode the specific timing behavior of each
commercial IC and the custom designed pH sensor chip in the FPGA modules as “controllers”, while exposing the handle to the actual value of all main operating parameters into the software domain. A simple example is the clock configuration. There are two clocks need to be generated, both of them are down-counted from the master clock running at 100 MHz. Their actual frequencies are specified on the software side, which are the threshold value that resets the counter. To ensure debuggability, the specification of the clock period is not only an one way write to the FPGA, but rather, a handshake operation involving bi-directional exchange of trigger signals. In this way, the FPGA side can update the clock counter in a way that avoids the creation of artifacts, which may corrupt the operation of IC’s connected to the clock signal, while letting the software keeps track of exactly when the change of clock frequency happens. It is specifically useful in this case, where the driving clock for the pH sensor chip is running at an extremely low speed. Similar design methodology is used for virtually all control signals.

A block level diagram of the Verilog modules used in the FPGA is shown in Figure 4.21. Since the complexity of the design remain fairly low, the utilization of the FPGA’s resources is not a concern, and the discussion on the actual implementation will not be extended further here.

Since the total data rate is dominated by the ADC, by down tuning the sample rate of the ADC to 10 kSamples/s, the total data traffic can be reduced to approximately 15 kB/s, well within the limit of what Opal Kelly’s front panel USB 2.0 interface can support (several MB/s, depending on the mode of data transmission). Under this scenario, the data transmission on the FPGA and data recording on the PC can be easily implemented under the assumption of no possible data traffic congestions.

The software is written using PyQt5 framework, with a mixed usage of Python and TOML scripting language. Multithreading is used to keep a smooth user interface (UI) experience, as the polling of different types of data input from the Opal Kelly is not pushed to the background. A drawing depicting the relationships between different threads is available as Figure 4.22. A trigger manager thread (red) governs the interaction between the software and the FPGA. Almost all events on the hardware side have handshake triggers, data ready triggers and exception triggers.
Asynchronous event objects are used to propagate the responses from the hardware side triggers to different threads, while the two main data threads (one for ADC output and one for pH acquisition from the chip) and pulls the data only when certain events are found to be set. Finally, the main UI thread distributes the user inputs to different threads, and pull values from the ADC and pH data synchronization threads through queues to update what will be displayed. It also automatically records the output into the memory, and saves the data to the file system on a regular basis.

A screenshot of the user interface is shown in Figure 4.23, where the top of the software window is used to change the clock frequency used in the pH sensor, the left panel is used to change the current bias for each of the two chips, and the recorded results are displayed in real time in the center.

4.3.4 pH Measurement

To verify the performance of the designed pH sensor, 5 different buffered pH solutions are prepared. The composition of the pH buffers is documented as follows:
Figure 4.22: Relations between different threads used in the software.

1. Citric buffer (10 mM, pH = 6.00): citric acid (C₆H₈O₇, Sigma-Aldrich) and sodium citrate dihydrate (Na₃C₆H₅O₇, Sigma-Aldrich Fine Chemicals), titrated using 0.1 M citric acid solution.

2. MOPS buffer (10 mM, pH = 7.00): MOPS (C₇H₁₅NO₄S, ACROS) and EDTA (C₁₀H₁₆N₂O₈, Sigma-Aldrich), titrated with 0.1M hydrochloride acid (HCl).

3. Tris buffer (10 mM, pH = 8.00 and pH = 9.00): tris hydrochloride (C₄H₁₁NO₃·HCl, Fisher Scientific) and tris base (C₄H₁₁NO₃, Roche Diagnostic GmbH), titrated using 0.5 M tris base solution.

4. Carbonate buffer (10 mM, pH = 10.00): sodium bicarbonate (NaCO₃, Sigma), and sodium
carbonate anhydrous (Na$_2$CO$_3$, Fisher), titrated using 0.5M sodium carbonate solution.

A commercial handheld pH sensor (Mettler Tolado FG2 with LE438 pH electrode) is used to calibrate the pH in the buffer solutions during preparation. With a 3-point calibration using commercially available buffered pH solutions at pH=4.00, pH=7.00, and pH=10.00, this handheld pH sensor is capable of delivering readouts at ± 0.05 pH accuracy [108].

The response of the IC against pH changes around the sensing site can be fully characterized, once the pH buffer solutions are ready, the IC is post fabricated and wirebonded, and the PCB-based testing system is fully assembled. To do this, an external silver-silver chloride (Ag/AgCl) QRE is attached to the $V_{DDA}$ of the NMOS ISFET version of the standalone packaged IC. 1 mL of the buffer solution of interest is then dropped onto the sensing area with a pipette, and in touch with the Ag/AgCl QRE. An 800 mV supply voltage is used to power the integrated pH sensor chip, and a 25 Hz clock is supplied externally to generate 1 sample/s pH readout. Digital output from the chip is then recorded on the computer, and its readout values are back assigned to the pH of the buffer solutions in use.

Figure 4.24 shows the pH response of the sensor from pH=6.00 to pH=10.00. In the figure, the last 240 seconds of the data are shown, with the last 30 seconds of the data as insets before swapping
Figure 4.24: The time domain behavior of the pH sensor between pH=6.00 to pH=10.00.

to a different pH solution. It is easy to observe the stable and relatively low noise data output from the designed integrated pH sensor. Unfortunately, there is no built-in testability to better quantify this. The lack of testability comes from the worry that the added parasitics from the test structure may degrade the performance. Thus it is impossible to have a separate characterization of each step in the analog signal chain to isolate the distortion and noise introduced from each stage.

Another issue to be noted, is that Tris-based buffers exhibit a much larger time constant before the readout value becomes stable. This is possibly because of a slow surface kinetics between the silicon dioxide and the counter ions in the Tris & Tris-HCl buffer solution at a relatively high pH. A slower response is consistently observed in ISFETs when pH values are higher [109], and the effect from different counter ions present in the solution has not been well studied at the time of writing [110].

The experiment is repeated until at least 3 data points are collected for each pH value from one packaged chip. This result is plotted in Figure 4.25. The end-to-end sensitivity is derived from the best linear fit of the acquired data points. The best estimated sensitivity is 65.8 LSB/pH between
Figure 4.25: End-to-end sensitivity of the pH sensor.

the source pH difference and the end digital output, a value corresponds to the slope of the fitted curve. This translates to a 25.7 mV/pH sensitivity in the front end, assuming negligible gain error (switched capacitor gain = 2 V/V) in the analog signal chain. Since this only achieves about 43% of the Nernstian limit, further post processing with modified sensing layer, as well as an etched down top layer passivation for less floating gate capacitive attenuation may further improve the sensitivity, and in turn improves the accuracy of the designed pH sensor.

Figure 4.26: 5-hour long measurement demonstrating the stability of the sensor.
Finally, the stability of the sensor is measured with the sensing site submerged in MOPS pH = 7.00 buffer for a 5-hour long soak. The readout during this experiment is plotted in Figure 4.26 in gray. Since the plot captures both long-term noise performance as well as the drift, two observations can be made from this recorded data set. The first observation is that the noise level is mostly within ± 1 LSB. A higher than 9 effective number of bits (ENOB) performance is likely accomplished in the ADC for noise alone, however, no evidence supports this hypothesis as the ADC cannot be tested separately. If a 4 mHz corner frequency, second order low pass filter is applied to the dataset, the long term drift is now exposed, and plotted in black. Over the 5-hour recording time, a maximum of 7 LSB drift is observed, which is the dominant factor degrading the sensitivity of the integrated pH sensor system.

4.3.5 Electrical Measurement

Finally, extra electrical measurements are done to characterize the power consumption of the pH sensor. It turns out that the assembled PCB system is not very helpful, since the majority of the power is dissipated in charging the capacitor loading the data output pin. It is easily observed since current spikes saturates the output when “0” to “1” data transitions happen. The power consumption from the load capacitor should not be counted as a part of the power consumed by the pH sensor IC, as it is the user’s responsibility to provide a proper load. For instance, in the integrated version, the load capacitor is at femto-Farad level, providing a negligible power overhead. With well encapsulated sensor ICs, it is hard to disconnect the wirebond underneath the protection epoxy. Thus power consumption is characterized using probe station with bare dies instead.

Figure 4.27 shows the transient response of the IC measured under the probe station, under a supply voltage of 800 mV, a clock frequency of 25 Hz, and a current code of 0x20. This is the exact condition for all previous pH related measurements. With the functionality of the chip verified, the data output pin is then disconnected from the probe station for supply current recording. An average current level of 0.9 nA (repeated measurements indicates an accuracy no higher than 0.1 nA) is measured, leading to a 0.72 nW power consumption.
Finally, to find out how wide the fully integrated pH sensor can be tuned in the performance space, a sweep in the clock frequency is performed at the same power condition. It turns out that the pH sensor can deliver 2 samples/s while maintaining sub-nW operation. With a change in the bias current, the pH sensor can trade maximum sampling frequency for a lower power consumption. These results are summarized and plotted in Figure 4.28.

4.4 Summary

In this section, a sub-nW power consumption pH sensor is presented as an example of a potential bio-sensing module to expand the prototyped ultrasonography compatible distributed sensor system presented in Chapter 3. Although several design choices are taken with limitations from what the ultrasound power and data frontend can provide, the fully integrated sensor’s performance is still outstanding among published ISFET based pH sensor ICs. Table 4.4 shows a comparison between this work and several low-power digital ISFET based pH sensor systems. It can be seen from
comparison that, overall, this work achieves the best output sensitivity, with the lowest power ever reported at a reasonable speed. It also consumes the smallest amount of area. Noticing the fact that this design uses the I/O MOSFETs almost exclusively (in fact, only the current source uses core devices), the equivalent technology node of this IC is closer to a 0.35 \( \mu \)m one. Yet the smaller area achieved in this work comes from the massive simplified frontend topology, especially compared to a similar chip designed in 0.18 \( \mu \)m presented in [96].

The FoM in this case, is defined in a way similar to that of Walden’s figure-of-merit for ADCs [102], as our best effort to capture the energy spent per bit accuracy acquired in pH sensing. However, digital sensitivity of the sensor is used instead of accuracy at the output. This is because among all publications that the author has surveyed, none of them reports the accuracy in the paper. Sensitivity is equivalent to one over accuracy, if the output of the pH sensor is accurate to the LSB. This is never the case, however, as at least a close to 1-bit resolution degradation will happen for any practical ADCs. For this work, the key factor that reduces the accuracy is the long term drift, pushing the confidence to assign a specific output code back to a pH value down to around \( \pm 0.1 \) pH. Yet it still outperforms the two digital pH sensors reported above. With power taken into consideration, this work provides a 4000 \( \times \) improvement in the FoM.

On the other hand, this designed pH sensor is surely not flawless. Even from the comparison
<table>
<thead>
<tr>
<th>Technology</th>
<th>0.18 µm</th>
<th>0.18 µm</th>
<th>0.6 µm</th>
<th>0.35 µm</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISFET Frontend Topology</td>
<td>Pseudo-Differential Source Follower</td>
<td>Differential CVCC with CDAC</td>
<td>Differential CVCC</td>
<td>Chemical Gilbert Cell</td>
</tr>
<tr>
<td>Digital Output</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Sensitivity (Output, Unit Varies)</td>
<td>65.8 LSB/pH</td>
<td>7.86 LSB/pH</td>
<td>37 LSB/pH</td>
<td>5.5 nA/pH</td>
</tr>
<tr>
<td>Sensitivity (Source, mV/pH)</td>
<td>25.7</td>
<td>9.7</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>Area (mm²)</td>
<td>0.85</td>
<td>1.45</td>
<td>21.4</td>
<td>8.01</td>
</tr>
<tr>
<td>Power Consumption (nW)</td>
<td>0.72</td>
<td>176</td>
<td>29.7k</td>
<td>165</td>
</tr>
<tr>
<td>Sampling Frequency (Sample/s)</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>FoM: Power/(F₅×Sensitivity) (nW·s·pH/LSB)</td>
<td>0.011</td>
<td>44.8</td>
<td>802.7k</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 4.4: Comparison between this work and other digital pH sensor systems, or low power pH sensor systems.

chart, it is obvious that there is still plenty of room to improve in the source sensitivity. This can be achieved through a better ion-sensitive layer deposited on the sensing area.

It is also worth noting that ion sensors, like [113], that directly buffers the Nernstian voltage on the sensing layer exists with an even lower power consumption reported. Yet the key advantage of using the ISFET/REFET pair remains in the fact that a well matched local reference can be generated reliably at a low cost, thanks to the rapid advancement of the IC technology.
Chapter 5: Conclusion

This thesis presents the design of the world’s first ultrasound sonography compatible implant sensor platform to the author’s best knowledge, together with a sub-nW integrated pH sensor that is designed to be compatible with the platform, for truly distributed, miniaturized, parallel operating deep-tissue bio-sensing applications. However, the author’s contribution to the field of bioelectronics is not limited, and will not be limited to the ideas presented in this work. This chapter will review the list of contributions of the author as a Ph.D. candidate, followed by some known limitations in the work presented in previous chapters, as well as some potential improvements and exciting novel ideas that can potentially lead to a newer, and/or further optimized design than the prototype works presented in this thesis. Finally, the concluding remarks is presented at the end of this chapter.

5.1 Summary of Contribution

On top of ultrasound sonography compatible sensor platform initially presented at [44], and the sub-nW pH sensor initially presented at [76], the author has also dedicated a significant amount of time on the following projects.

The first project is an electrochemical camera chip for real time, high spatial resolution redox molecule imaging [114], which is also the first project that I got involved in as a graduate student. An application of this work has helped to promote the understanding of the interaction between the redox-active phenazine 5-Me-PCA and the efflux pump MexGHI-OpmD in the gram-negative pathogen Pseudomonas aeruginosa [115]. A separate electrochemical stimulation platform was later built to further investigate the interaction between an externally supplied oxidative potential and the morphology in a Pseudomonas aeruginosa biofilm [116], as a direct verification of the
phenazine’s role as an electron shuttle.

Finally, a passive rectifier design work is integrated into [117], as a part of a miniaturized, implantable, ultrasound powered temperature sensor.

5.2 Future Work

The work presented here is by no means flawless, and even if they are optimized, there are still plenty of novel ways to make them even better. Several such cases are discussed here, including what can be done in the future on the circuit side to potentially take the ultrasonography compatible sensor system to the next level.

5.2.1 The Ultrasonography Compatible Sensor Frontend

The ultrasound imaging compatible power and data telemetry system, as a pioneering prototype to demonstrate the possibility, is designed in a conservative way to minimize the risk anticipated in the testing. Yet after the ICs are fabricated, the measurement result has indicated that at least a great amount of performance margin exists power-wise. This can be used to implement more complicated logic functions, such that a smarter clock and data recovery scheme can be implemented, leading to a more robust operation. Also, the application layer will have a plenty of room to expand.

On the other hand, there are two techniques that the author believes a potential implementation in the sonography compatible sensor system would bring this work to the next level. The first is a potential usage of the harmonic backscatter [118]. Unlike the backscatter mechanism implemented in this work, harmonic backscatter modulates the third (or higher) order component in the backscattered wave. It can potentially decouple the data uplink with power, clock recovery, and data downlink, if this technique can be translated to designs working with heavily duty cycled ultrasound pulses. With this, duplex data communication becomes possible, which has the potential to increase the data bandwidth by 3 times.

The second is a potential usage of a tunable, adaptive matching network to take advantage of the inductive band shown in Figure 3.7 [119]. The efficiency of a switch-only (or other switch-mode
harvester) at the moment is still much lower than what a conjugate matched power delivery can achieve. A pulsed ultrasound compatible version of this will drastically increase the power budget of the sensor system, opening up possibilities for much more complicated functionalities.

5.2.2 The Sub-nW Integrated pH Sensor

Even as a stand alone pH sensor, the designed IC still possesses massive potential to be further optimized.

The first optimization is in fact a correction, which comes from mistakes made in the calculation when designing the signal amplification chain. Together with a better understanding on the model inaccuracy, at least a 300 pW of extra power saving should be possible.

The second one is a limited life time. This is also what the author sees as the biggest shortcoming of this work. The dominant life time degradation comes from electrochemical etching of the copper traces underneath the passivation, and this has been observed for all chips after 48 hours of continuous experiment. One of the reasons for this to happen is that the manual encapsulation process cannot have a μm-level accuracy, to expose only the sensing site. This problem can be easily alleviated by a modification to the post-fabrication process, to add a “passivation helper layer” on top of the IC. Right now electrochemical etching is the dominant life-time-defining mechanism, however, once this is resolved, there are several other candidates that may become the bottleneck for stability in long-term performance.

The last problem the author would like to discuss here has hints spread across Chapter 4. In Subsection 4.2.2, it is stated that with an ISFET/REFET pair, the variation on the reference potential will translate to a common mode variation, orthogonal to the differential signal. Yet in Subsection 4.2.3, Equation 4.9, what was effectively described is that the gain and the distortion are both functions of the common mode voltage. Though this CMRR could introduce only negligible degradation to the signal quality of the final output, simple circuit tricks exists to massively reduce this common mode voltage variation at a potentially low cost, yet a sub-nW version may still requires some novel work.
5.3 Final Remarks

During the 2020 Symposia on VLSI Technology & Circuits, foundries around the world share their newest development on the next generation of CMOS devices. Sub-7 nm FinFET, stacked nano sheet with gate all around, 3D-integration, all these technologies seem promising to keep Moore’s prediction well held for another 5 years onward. Yet if people were standing at year 2000, when 180 nm bulk CMOS was just widely commercialized, would they believe that in 20 years, the industry can massively fabricate FinFETs at 7 nm, that powers up such an enormous amount of digital services all around the world? But for people who already stand at the 7 nm node right now, it might be obvious for them to see what potential is still left there for a further down scaling.

Similarly, looking back at the time before I delved into biomedical IC design, it would be impossible for me to tell how a better biomedical implant can be designed. Yet only after iterations of IC designs do I start to realize that not only how sub-optimal they are, but also without them, I would not have ever figured out what to improve. The conclusion of this thesis is by no means the conclusion of my research life, but rather the starting point to explore the world of finer designed ICs for implantable bio-sensing applications.

People who reach close to the end of the “more Moore” highway have just shown the world where they can drive to next. Yet I hope the work presented in this thesis has demonstrated what one of the rocky paths under the name of “more than Moore” can reach, and where it may lead to in the near future.
References


peripheral nerves,” *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 2, pp. 257–270, 2018.


