An Ultra Low Power Implantable Neural Recording System for Brain-Machine Interfaces

by

Woradorn Wattanapanitch

B.S., Cornell University (2005)
S.M., Massachusetts Institute of Technology (2007)

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

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author .....
Department of Electrical Engineering and Computer Science
May 18, 2011

Certified by .....
Rahul Sarpeshkar, Ph.D.
Associate Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski, Ph.D.
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

An Ultra Low Power Implantable Neural Recording System for Brain-Machine Interfaces

by

Woradorn Wattanapanitch

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

In the past few decades, direct recordings from different areas of the brain have enabled scientists to gradually understand and unlock the secrets of neural coding. This scientific advancement has shown great promise for successful development of practical brain-machine interfaces (BMIs) to restore lost body functions to patients with disorders in the central nervous system. Practical BMIs require the uses of implantable wireless neural recording systems to record and process neural signals, before transmitting neural information wirelessly to an external device, while avoiding the risk of infection due to through-skin connections. The implantability requirement poses major constraints on the size and total power consumption of the neural recording system.

This thesis presents the design of an ultra-low-power implantable wireless neural recording system for use in brain-machine interfaces. The system is capable of amplifying and digitizing neural signals from 32 recording electrodes, and processing the digitized neural data before transmitting the neural information wirelessly to a receiver at a data rate of 2.5 Mbps. By combining state-of-the-art custom ASICs, a commercially-available FPGA, and discrete components, the system achieves excellent energy efficiency, while still offering design flexibility during the system development phase. The system's power consumption of 6.4 mW from a 3.6-V supply at a wireless output data rate of 2.5 Mbps makes it the most energy-efficient implantable wireless neural recording system reported to date. The system is integrated on a flexible PCB platform with dimensions of 1.8 cm $\times$ 5.6 cm and is designed to be powered by an implantable Li-ion battery.

As part of this thesis, I describe the design of low-power integrated circuits (ICs) for amplification and digitization of the neural signals, including a neural amplifier and a 32-channel neural recording IC. Low-power low-noise design techniques are utilized in the design of the neural amplifier such that it achieves a noise efficiency factor (NEF) of 2.67, which is close to the theoretical limit determined by physics. The neural recording IC consists of neural amplifiers, analog multiplexers, ADCs, serial programming interfaces, and a digital processing unit. It can amplify and

digitize neural signals from 32 recording electrodes, with a sampling rate of 31.25 kS/s per channel, and send the digitized data off-chip for further processing. The IC was successfully tested in an *in-vivo* wireless recording experiment from a behaving primate with an average power dissipation per channel of 10.1 μW. Such a system is also widely useful in implantable brain-machine interfaces for the blind and paralyzed, and in cochlea implants for the deaf.

Thesis Supervisor: Rahul Sarpeshkar, Ph.D.
Title: Associate Professor of Electrical Engineering

Acknowledgments

First of all, I would like to thank my advisor, Prof. Rahul Sarpeshkar, for his support and encouragement throughout my graduate school years. In these past six years under his guidance, I always regard him as an excellent example of a person with great technical brilliance, enthusiasm for learning, intuition to understand problems deeply, and, most of all, unwavering courage to tackle any difficult problem, which always inspires me to strive for the same. He not only has instilled in me technical knowledge for my future professional development, but also has taught me how to not give up when everything looked grim, and always encouraged me to try my hardest at overcoming all the obstacles that I am facing. I am truly grateful and privileged to be able to go through my graduate school years at MIT under his supervision.

What makes my graduate school experience truly enjoyable is the support from my colleagues including the past and present members of the AVBS group. My special thanks go to Dr. Soumyajit Mandal, Scott K. Arfin, Benjamin I. Rapoport, Daniel Kumar, Bruno Do Valle, and Dr. Lorenzo Turicchia. It has been an unforgettable experience to be able to work side-by-side with them throughout most of my years at MIT.

I am truly grateful for the support and encouragement of my family in Thailand. They always encourage me to focus on my study and accomplish my goal, and never worry about the situation at home. They have been waiting, day and night, for my return to Thailand in the past 11 years of my study in the United States, without ever complaining that it takes so long. I am eternally grateful to my father, Somsak Wattanapanitch, who instilled in me the love of science at a very young age. He always believed that one day his son would accomplish great things, and that belief has always encouraged me to never fail his expectation. I believe that his soul in heaven will be delighted to see that one of his son's dreams is about to be fulfilled.

Finally, I would like to give a special thank to my fiancée, Dr. Methichit Chayosumrit, whose love, support, and encouragement always make every single day of mine truly meaningful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter of the thesis motivates the need for ultra-low-power implantable wireless neural recording systems, informs the readers of prior works in this area, presents the design approach of the system to be discussed later in this thesis, and outlines the organization of the thesis.

1.1 Motivation

In the past few decades, direct recordings from the cortical area of the brain have enabled scientists to gradually understand and unlock the secrets of neural coding. With the aid of high-density microelectrode arrays, neural activities from a large population of neurons can be observed simultaneously with a spatial resolution down to that of a single cell [29], [35]. Many experiments in non-human primates [70], [64], [58] and a pilot clinical trial in a human subject [27] illustrated that control signals directly derived from spiking activities from a population of neurons in the cortical area of the brain can be used to successfully control and manipulate computer devices or robotic limbs. The study in [67] shows that cortical activities from a population of neurons can be used to control even a sophisticated device such as a robotic limb with multiple degrees of freedom. These studies have shown great promises for a successful development of practical brain-machine interfaces (BMIs) to restore lost body functions to patients with disorders in central nervous system

such as those suffering from spinal cord injuries. Practical BMI systems of the future will be portable and may enable the users to control dexterous robotic limbs or their natural limbs at a near-natural level.

Nevertheless, to reach a stage where such BMIs can be used chronically in humans, many challenges need to be solved. BMI systems require the use of neural recording systems to obtain neural data. In the studies mentioned earlier, neural recordings were performed with passive microprobes that were implanted in the cortical area of the brain. The implanted microprobes were then connected by a bundle of transcutaneous wires to external recording electronics. These recording electronics were normally large in size and consumed watts of power, and thus needed to be mounted on a rack or a subject's wheel chair. In addition, the transcutaneous connections also pose a major risk of infection due to the skin rupture, and thus must be eliminated in practical BMIs. Therefore, for clinically viable BMIs, the recording systems should be entirely implanted under the skin, while the recorded neural data and the power to operate the implants must be transferred through wireless means. This implantability requirement poses major constraints on the size and total power consumption of the recording systems.

To avoid excessive heat dissipation that may cause cell deaths in the surrounding tissues, the total power dissipation from the recording systems should be kept below a 10 mW range. For the battery-operated recording systems, low power consumption could prolong the time between recharges, thus expanding battery life to avoid frequent surgeries for battery replacements. Power dissipation of an implantable recording system is also a strong determinant of its size and cost. Low power consumption means that small batteries can be used to power the recording systems. In the case when the recording systems are to be continuously powered, a smaller RF coil can be used to receive RF power to operate the systems. With a small-sized implant, the cost of packaging reduces as well and the complexity of the surgical procedure for implanting such system may significantly decrease [53]. Due to these reasons, minimizing power consumption of implanted neural recording systems should be the priority in the design of BMIs.

1.2 Previous Work

Many recording systems with intended use in wireless neural recording applications have been reported in the literature. A number of design approaches have been pursued that portray the compromise among design flexibility, turnaround time, power consumption, and sizes. Advances in integrated circuit (IC) technologies have enabled engineers to increase the number of recording channels and signal processing functions that can be put on a single chip, while still decreasing the size and improving performance. Such technology advancement provides an excellent mean for a development of neural recording systems since a large number of recording channels, signal processing functionalities, and wireless communication circuitry can be integrated in a small form factor, while consuming low enough power to make full implantations of such systems feasible. Thus, for the purposes of minimizing power consumption and reducing the size of the systems, an application-specific integrated circuit (ASIC) approach in which all functionalities are custom designed as integrated circuits (ICs) is normally pursued. However, this ASIC approach has a slow turnaround time due to the design and fabrication of the ICs. Furthermore, design flexibility is sacrificed since the functionalities cannot be easily changed once the ICs have been fabricated.

Examples of the systems that utilized the ASIC approach were reported in [42], [26], [62]. The system reported in [42] contains a total of 32 neural recording channels, which are grouped into four neural probes. Each 8-channel neural probe contains a front-end selection circuitry that multiplexes from 64 recording sites to eight neural amplifiers on the neural probe. The outputs of the eight neural amplifiers on each neural probe are time-multiplexed to drive an ADC on a data-compression ASIC. The data compression ASIC then utilizes a window thresholding method for spike detection to reduce the amount of data that needs to be transmitted wirelessly. The spike waveforms that cross predefined threshold levels and the corresponding addresses of the electrodes are reported at 5-bit resolution. Without a wireless transmission feature, the total power of the recording system including the four neural probes and the data-compression ASIC was reported to be 5.4 mW. This multi-chip system is

integrated on a 3-dimensional platform. The system in [26] integrates all the functionalities including neural signal amplification, data reduction, neural signal digitization, and wireless communication into a single chip. It contains 100 neural recording channels and includes a wireless data transmission feature by a fully-integrated FSK transmitter. The power and commands are transferred from an external unit to the implanted system via an inductive power link. The system utilizes a simple thresholding scheme with analog spike detection circuitry to reduce the amount of data that needs to be transmitted. The system allows one raw analog channel to be selected for full digitization at 10-bit resolution by an on-chip ADC. The total power consumption of the system is 13.5 mW. Another system reported in [62] contains 64 recording channels. To record from 64 channels, the system utilizes four 16-channel neural preconditioning ASICs in parallel. The preconditioning ASICs are interfaced with a neural processing unit which consists of two 32-channel neural processing chips [61]. The system also contains a bi-directional telemetry chip for transmitting neural data to the external world, and for receiving power, commands, and clock to operate the implant. The overall multi-chip system is integrated on a Si-platform and consumes a total power of 14.4 mW.

At the other end of the spectrum, many recording systems are constructed from commercially available parts which are integrated into a system at the printed-circuit-board (PCB) level. Some of these systems contain programmable logic devices such as Field-Programmable Gate Arrays (FPGAs), complex programmable logic devices (CPLDs), or microcontrollers, to implement digital signal processing functions for the recording systems. Clearly, this approach results in a faster turnaround time compared to the full-ASIC approach, and offers greater design flexibility due to the uses of programmable logic devices or microcontrollers. With such approach, signal processing algorithms can be readily modified even after the hardware has been built. The clear disadvantage of this approach is its high power consumption and the larger size of the systems. As a result, most systems constructed from this approach are not yet suitable for chronic implantation in human subjects since the power consumption is still too high to be considered safe for the surrounding tissues. However, these

systems have proved to be tremendously useful in neuroscience studies where animal subjects are employed [52], [49]. The system in [52] consists of an analog module interfaced with a digital module. The analog module consists of two 8:1 input analog multiplexers that multiplexes 16 input channels into two neural recording channels. Each recording channel consists of a unity-gain buffer, a differential amplifier, and a filter, all built from commercially-available parts. The digital module consists of a microcontroller with a built-in ADC. The microcontroller is interfaced with a compact flash memory for storing neural data during the experiment. The system reported in [49] contains a total of 96 recording channels which are grouped into three digitizing headstage modules. Each 32-channel module consists of two custom ICs, with each IC containing 16 neural amplifiers, a 16:1 analog multiplexer, and a digital-to-analog converter (DAC) to control offset voltages of the amplifiers. The two custom ICs are interfaced with commercially-available ADCs, while the clock and control signals for the custom ICs are generated from a commercially-available CPLD. The three digitizing headstage modules are interfaced with the implantable central communication module. The communication module consists of an FPGA, which is designed to reduce the amount of data received from the three headstages [50], the RF data transceiver for data and commands communications, voltage rectifiers and power regulators to generate DC supply voltages from received RF power. The total power consumption of the implantable part of this system is close to 2 W.

1.3 Our Approach

It is my belief that, at present, the scientific community is still in an early stage of BMI system development. Neural data processing algorithms such as spike detection [40], [74], [9], spike sorting [73], [66], and neural data compression [39] are currently being developed by many research groups around the world to help improve performance of BMIs. While achieving low power consumption is a crucial aspect for the design of an implantable neural recording system, design flexibility should not be completely ignored. While an all-ASIC approach can result in low power consump-

tion and small form factor, at this stage, it might be too early for such systems to be widely useful. For instance, while the simple thresholding method in [26] might be effective at reducing the amount of data that needs to be transmitted, some important information such as spike amplitudes and spike shapes are lost, thus preventing the uses of many processing algorithms previously mentioned. On the contrary, flexible systems such as [49] can provide richer functionalities and these functionalities can even be modified during the experimental stage, even after the hardware has been built, by reprogramming the in-system programable logic devices. However, the total power consumption of close to 2 watts in [49] would prevent such system to be used chronically in human subjects.

The most suitable approach at the current stage of BMI development might lie somewhere in between these two extremes. By combining good energy efficiency from low-power ASICs, and design flexibility from commercial programmable logic devices, a low-power neural recording system that is also highly programmable can be built. This thesis presents a development of such neural recording system with the goal of practical use in brain-machine interfaces. The ASIC approach is utilized for parts of the system that require excellent energy efficiency, while an FPGA is used where design flexibility is more important. Figure 1-1 shows the conceptual diagram of the implantable wireless neural recording system to be presented in this thesis. The system consists of an internal unit and an external unit. The internal unit consists of a front-end processing stage that amplifies and digitizes neural signals from recording electrodes. The digitized neural data from the front-end processing stage is then processed by a digital signal processing module on the internal unit, before the processed neural data is transmitted to the external unit via a wireless data telemetry system. The external unit receives the neural data and relays it to a remote device such as a computer or a robotic limb. For system programmability, the communication between the internal unit and the external unit is bidirectional. In addition to transmitting the processed neural data from the internal unit to the external unit (uplink), the wireless data telemetry system can transmit commands in the reverse direction to configure the parameters of the internal unit (downlink). In

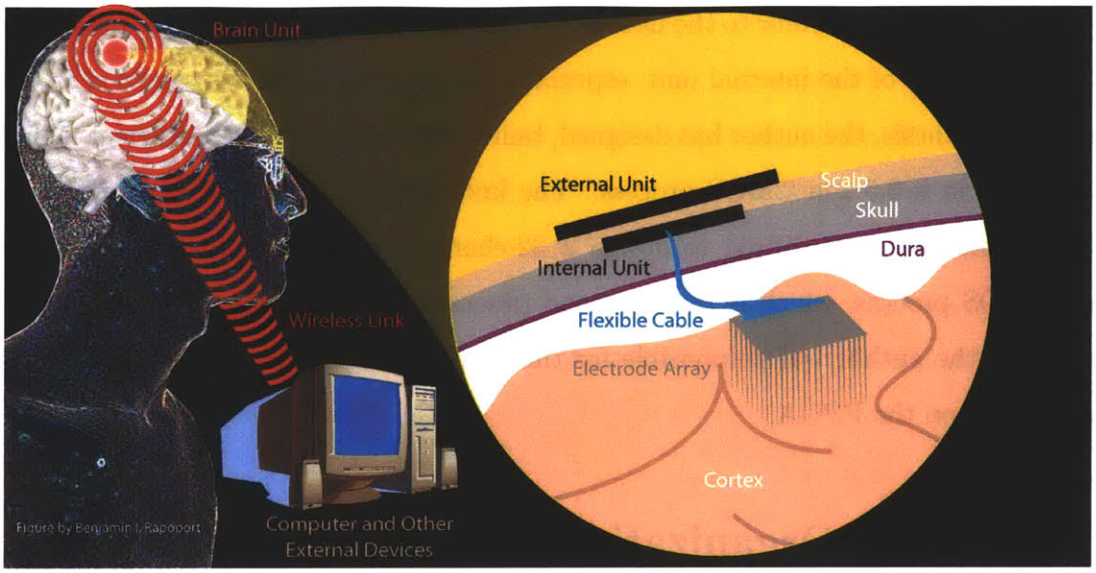


Figure 1-1: Conceptual diagram of our BMI system.

addition, the external unit is responsible for delivering power wirelessly to charge the implanted battery that powers the internal unit.

Due to the stringent requirement on power dissipation inside the body, the goal of this thesis is to minimize power dissipation of the internal unit without compromising its performance. Since the power consumption of the front-end processing stage normally constitutes the majority of the total system power, it is implemented with a full-ASIC approach to achieve minimal power consumption. The processing of the digitized neural data is performed in an on-board FPGA to offer design flexibility during the system development.

1.4 Author's Contribution

It might be clear to the readers that the amount of work required to develope an implantable wireless neural recording system, such as that shown in Fig. 1-1, is beyond what one PhD student can handle. The author has been very fortunate to be a part of an excellent research team, and to be the one responsible for integrating the system from various subsystems that other team members have designed. The major

contributions of the author to the neural recording system development are the design and integration of the internal unit, especially the signal processing aspects of it. As part of this thesis, the author has designed, built, and tested an energy-efficient neural amplifiers in a 0.5 μm CMOS process. The knowledge gained from designing such neural amplifier was utilized to design a 32-channel neural recording IC in a 0.18 μm CMOS process, which is the front-end processing stage of the internal unit. In addition, the author was responsible for the design of signal processing and control algorithms on the FPGA.

1.5 Thesis Organization

While the overall neural recording system will be discussed, this thesis will focus on the design of the internal unit, especially on its signal processing and system control aspects. Chapter 2 presents the design, power and noise minimization techniques, and the experimental measurements of an energy-efficient neural recording amplifier. After the neural amplifier has been introduced, Chapter 3 presents the design and experimental measurements of the 32-channel neural recording IC that forms the heart of the internal unit of the neural recording system. The neural amplifier's topology of the system described in Chapter 3 is greatly influenced by the design techniques presented in Chapter 2. In Chapter 3, the design of each system component of the neural recording IC including the neural amplifier, the analog multiplexer, the analog-to-digital converter, the digital control unit, and the serial programming interface unit will be presented in detail. Experimental measurements from a wireless recording setup in a behaving non-human primate is also presented in this chapter. Chapter 4 discusses in more detail the design of the overall neural recording system, with the emphasis on the design of the internal unit and its signal processing aspect. Experimental measurements of the overall neural recording system is also presented. Chapter 5 concludes the thesis, discusses future work, and summarizes the author's contributions to the field of neural recording system design.

Chapter 2

An Energy-Efficient Micropower Neural Recording Amplifier

One of the most important components of BMIs is the neural signal amplifier. Neural signals from extracellular recording are very weak (typically with amplitude between 10 μV and 500 μV). As a result, amplification is needed before such signals can be processed further. Next generation high-channel-count BMIs will incorporate a large number of neural amplifiers (on the order of 100-1000, one for every electrode) to improve the decoding performance. For such applications, ultra-low-power operation is very important to minimize heat dissipation in the brain, preserve long-battery life, and maximize the time between recharges. To get clean neural signal recordings, it is important that the input-referred noise of the amplifier be kept low. Practically, the input-referred noise of the amplifier should be kept below the background noise of the recording site (5 μV-10 μV) [23]. However, designers must address the tradeoff between low-noise and low-power designs of the amplifier. For an ideal thermal-noise-limited amplifier with a constant bandwidth and supply voltage, the power of the amplifier scales as $1/v_n^2$ where v_n is the input-referred noise of the amplifier. This relationship shows a steep power cost of achieving low-noise performance in an amplifier.

Prior to our design being reported in [69], many designs of neural amplifiers had been reported in the literature [24, 36, 41, 45]. Most of these designs consume power

near 100 μW to achieve less than 10 μV_{rms} input-referred noise for bandwidths of 5-10 kHz. The designs in [36, 41] consume power near 100 μW to achieve about 8-9 μV_{rms} input-referred noise with approximately 10 kHz of bandwidth. The design in [24] achieves an input-referred noise of 2.2 μV_{rms} with 7.2 kHz of bandwidth while consuming 80 μW of power. If such amplifiers are to be used in a multi-electrode array, with a power near 100 μW per amplifier for most designs, the power required for the neural amplifiers can become the limiting factor for the whole multi-electrode system. To address this problem, we present a new micropower neural recording amplifier design. With our design, the power consumption per amplifier is low enough such that the total power consumption of a multi-electrode array may no longer be the bottleneck for the design of brain-machine interfaces.

This chapter is organized as follows. Section 2.1 discusses the high level operation of the amplifier. Section 2.2 describes the design principles and noise analysis used in the OTA to achieve a good power-noise tradeoff. Section 2.3 presents some measured lab bench and *in-vivo* results of the amplifier configured for neural spike recording. Section 2.4 presents experimental results when the amplifier is configured for LFP recording. Section 3.8 concludes the paper.

2.1 Overall Architecture of the Neural Amplifier

The overall schematic of the neural amplifier is shown in Fig 2-1. The topology of the gain stage is similar to the design in [24]. It uses a capacitively-coupled architecture to reject the DC offset that occurs at the electrode-tissue interface. This design includes a bandpass filter stage following the gain stage to shape the passband of the amplifier. The low-frequency high-pass cutoff of the gain stage is created by the MOS-bipolar pseudoresistor element [14] formed by $M_{b1} - M_{b2}$ and the capacitance C_f. The capacitive feedback formed by C_f and C_{in} sets the midband gain of the amplifier to approximately 40.8 dB. The high-pass cutoff and the low-pass cutoff frequencies of the amplifier can be adjusted via V_{tune} and the bias current of the g_m-OTA in the bandpass-filter stage respectively. With the addition of the bandpass-filter stage,

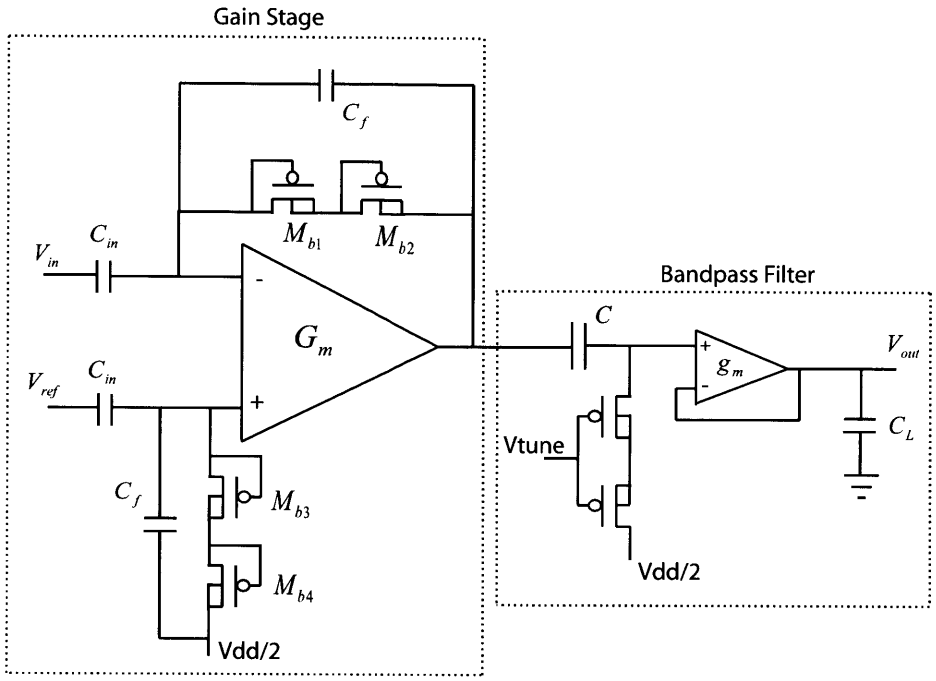


Figure 2-1: Overall architecture of the Neural Amplifier.

the amplifier can be configured to record either the local field potentials (LFPs) ($<$ 1 Hz to 300 Hz) or neural spikes (300 Hz to $>$ 1 kHz). For low-bandwidth LFP recording, the bias current of the OTA in the gain stage can be lowered to conserve power. It is worth mentioning that the high-pass cutoff frequency of the gain stage should be kept as low as possible. As reported in [45], placing a weak-inversion MOS transistor in parallel with C_f to create a high-pass filter with a cutoff frequency at a few hundred Hz introduces low-frequency noise that rolls off as $1/f^2$ in power units due to the noise from the transistor being low-pass filtered by C_f. This low-frequency noise appears at the front-end and gets amplified by the gain of the amplifier thereby degrading the minimum detectable signal. In our design as well as in [24], however, the MOS-bipolar pseudoresistor element's noise is at very low frequencies since the MOS-bipolar pseudoresistor element has a much higher impedance than a weak-inversion MOS transistor. Therefore, low-frequency noise due to this element is filtered out well before the passband and does not appear in the frequency band of interest.

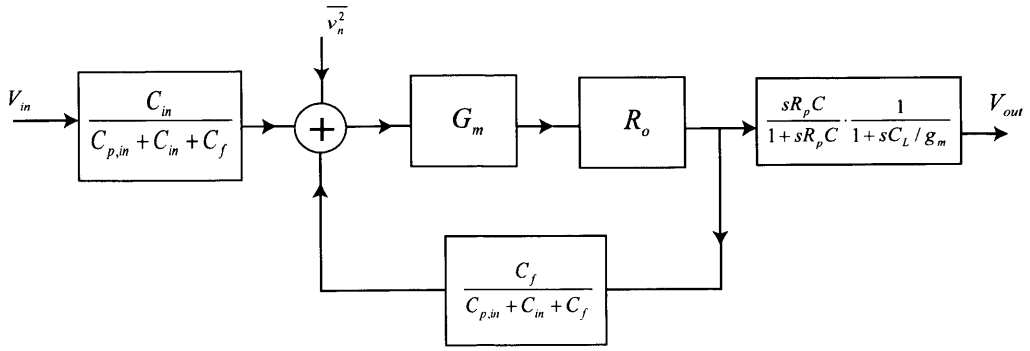


Figure 2-2: Block diagram of our neural amplifier including the input noise source of the OTA.

The operation of our amplifier can easily be understood by the block diagram of Fig 2-2. We include $C_{p,in}$ to model parasitic gate capacitances at input terminals of the gain-stage OTA. The input referred-noise of the OTA is modeled as a $\overline{v_n^2}$ term added to the system at the input of the gain-stage OTA. The gain-stage OTA is used as a high-gain amplifier and is modeled by G_m and R_o blocks where G_m and R_o represents the transconductance and the output resistance of the gain-stage OTA respectively. In the bandpass-filter stage, R_p is the resistance of the series PMOS transistors operating in the triode regime. The value of R_p is set by V_{tune}. The combination of C and R_p realizes the highpass cutoff frequency for the amplifier. From the small-signal block diagram in Fig. 2-2, assuming that G_mR_o is much higher than 1, we can express the transfer function of the neural amplifier as

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = -\frac{C_{in}}{C_f} \cdot \frac{sR_pC}{1+sR_pC} \cdot \frac{1}{1+sg_mC_L}. \tag{2.1}$$

The midband gain of the amplifier is $A_v = -C_{in}/C_f$. The highpass cutoff frequency is at $f_{HP} = 1/(2\pi R_pC)$ whereas the lowpass cutoff frequency is at $f_{LP} = g_m/(2\pi C_L)$. We can relate the input-referred noise $\overline{v_n^2}$ of the gain-stage OTA to the input-referred noise $\overline{v_{n,amp}^2}$ of the overall amplifier as

$$\overline{v_{n,amp}^2} = \left(\frac{C_{in}+C_f+C_{p,in}}{C_{in}}\right)^2 \cdot \overline{v_n^2}. \tag{2.2}$$

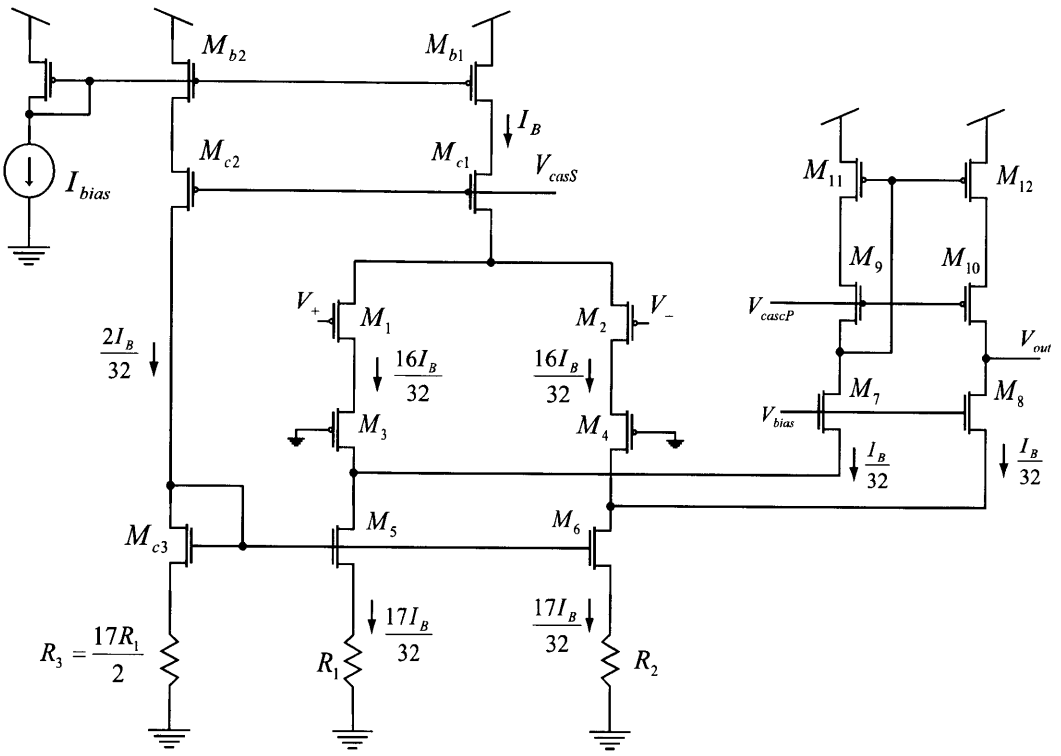


Figure 2-3: Schematic of the low-noise OTA used in this design.

The input-referred noise of the bandpass filter stage is insignificant and is not included in the block diagram since the gain of 40 dB of the gain stage alleviates the bandpass-filter stage's input-referred noise requirement. As a result, the power consumption of the bandpass filter stage is much smaller than that of the gain stage. Thus, to achieve low-noise performance, it is important to design the gain-stage OTA to have low input-referred noise. Section 2.2 describes the low-noise low-power design techniques used in this OTA.

2.2 Low-Power Low-Noise OTA design for gain stage

The schematic of the low-noise OTA is shown in Fig. 2-3. It is a modified version of a standard folded-cascode topology shown in Fig. 2-5. The OTA in Fig. 2-3 is biased such that the currents of the transistors in the folded branch $M_7 - M_{12}$ are only a small fraction of the current in the input differential pair transistors M_1 and M_2. In our design, the channel current in $M_7 - M_{12}$ is scaled to approximately $1/16^{th}$

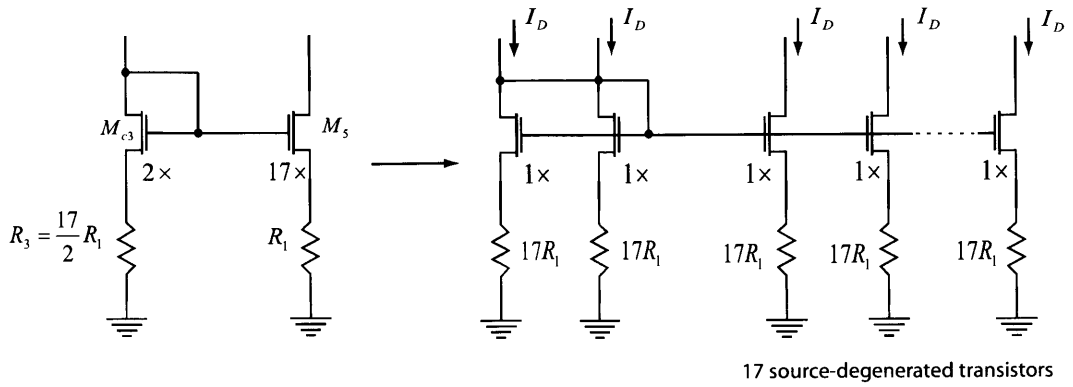


Figure 2-4: Circuit schematic for analyzing current scaling in the source-degenerated current mirrors of Fig. 2-3.

of the current in M_1 and M_2. The much lower current in $M_7 - M_{12}$ makes the noise contributed by them negligible compared to that from M_1 and M_2. As a result, we simultaneously lower the total current and the total input-referred noise of the OTA.

To ensure that such severe current scaling is achieved, we carefully set the bias currents of M_5 and M_6 through the use of the bias circuit formed by M_{b2}, M_{c2} and M_{c3}. The current sources M_{b1}, M_{b2} are cascoded to improve their output impedances and thereby ensure accurate current scaling. They operate in strong inversion to reduce the effect of threshold voltage variations. The source-degenerated current mirrors formed by M_{c3}, M_5 and M_6 and resistors R_1 and R_2 set the currents in M_5 and M_6 such that the currents in M_7 and M_8 (the difference between the current in M_3 and M_5 and between the current in M_4 and M_6) are a small fraction of the currents in M_1 and M_2. An analysis of mismatches in source-degenerated current mirrors is deferred until Section 2.2.3 and is important for robust biasing performance. In order to save power in the bias circuit, the current scaling ratio between M_{b1} and M_{b2} is 16:1 ($2I_B/32$) as shown in Fig. 2-3. To set the currents in the folded-branch transistors to be $I_B/32$, which is $1/16^{th}$ of the currents in differential-pair transistors, we set the current in M_5 and M_6 to be $17I_B/32$. Such current ratioing is achieved by making R_3 to be $17R_1/2 = 17R_2/2$, and constructing M_{c3} as a parallel combination of two unit transistors while M_5 and M_6 are each constructed from 17 unit transistors in parallel. To clarify this scaling further, the current mirror formed by M_{c3}, R_3

and M_5, R_1 in Fig. 2-3 is transformed into an equivalent circuit comprised of many source-degenerated unit transistors as shown in Fig. 2-4. All source-degenerated unit transistors are identical and have the same gate voltage. For any gate voltage there is only one source voltage at which a unit resistor's current equals a unit transistor's current. Thus, the nominal channel currents in all unit transistors are identical and the total current in M_5 is 17/2 times the current in M_{c3} as desired.

For the amplifier to have low input-referred noise, the transconductance G_m of the OTA needs to be maximal for a given current level. For the standard folded-cascode OTA shown in Fig. 2-5, the impedance looking into the sources of M_5 and M_6 is much smaller than the impedance looking into the drains of $M_1 - M_4$. As a result, the standard folded-cascode OTA achieves an overall transconductance G_m near g_{m1}, the g_m of M_1. However, if we lower the current in $M_5 - M_{10}$ to be a small fraction of the current in M_1 and M_2, the impedance looking into the sources of M_5 and M_6 can be a significant portion of the impedance looking into the drains of $M_1 - M_4$ such that incremental currents do not almost all go through the sources of M_5 and M_6 in the current divider formed between the sources of M_5 and the drains of M_1 and M_3. Therefore, G_m is significantly less than g_{m1}. Section 2.2.1 explains how we achieve G_m near g_{m1} even with our extreme current scaling via the use of source-degenerated transistors M_5 and M_6 in Fig. 2-3.

In the standard folded-cascode topology shown in Fig 2-5, the current sources formed by M_3 and M_4 contribute a significant amount of noise due to their large channel currents. In this design, we replace the current-source transistors M_3 and M_4 in Fig. 2-5 with source-degenerated current sources formed by M_5 and M_6 and degeneration resistors R_1 and R_2 as shown in Fig. 2-3. With an appropriate choice of degeneration resistance, the noise contributions from the source-degenerated current sources are mainly from the resistors and can be made much smaller than the noise contributions from MOS transistors operating at the same current level. Another benefit of using source-degenerated current sources is that the noise from resistors is mainly thermal noise while NMOS current sources contribute a large amount of 1/f noise unless they are made with very large area. As a result, the 1/f noise in

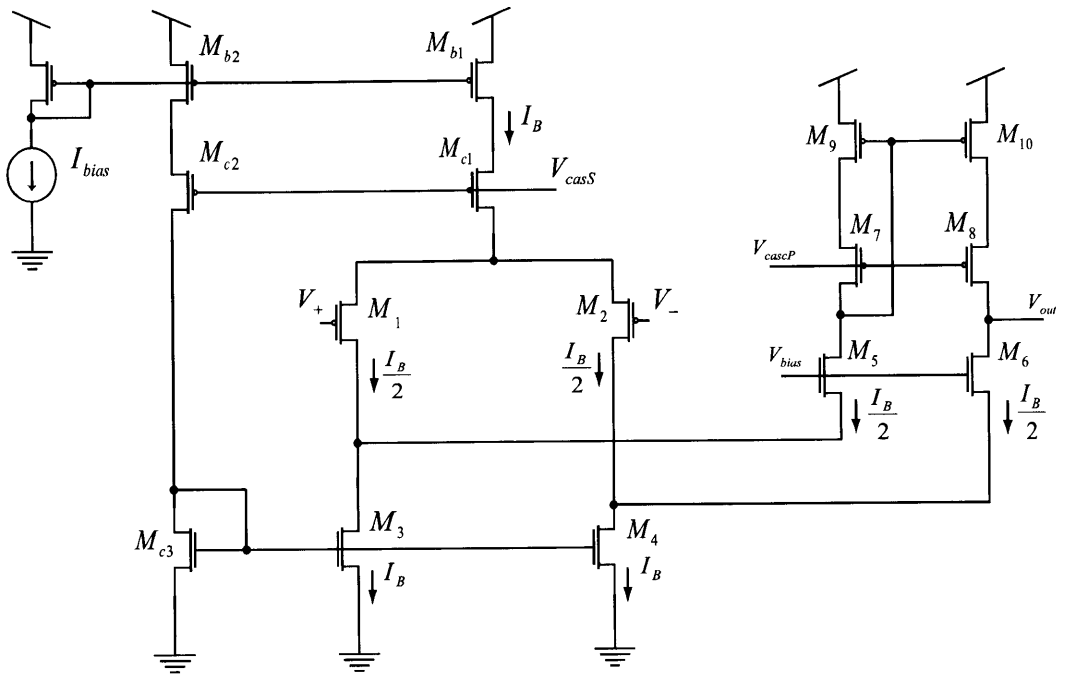


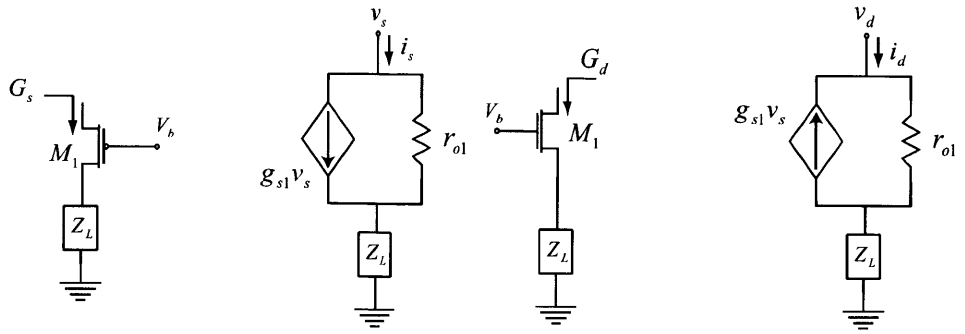
Figure 2-5: Schematic of a standard folded-cascode OTA.

our neural amplifier is mainly from the input differential pair. Therefore, the input-differential pair is made with large-area PMOS transistors, which have lower $1/f$ noise than similarly-sized NMOS transistors in most CMOS processes.

2.2.1 Device sizing for maximizing G_m

To achieve low input-referred noise, it is important that the transconductance of the OTA be maximized for a given total current. The maximum transconductance of the standard folded-cascode OTA that can be achieved is the transconductance of one of the transistors in the input-differential pair, say g_{m1}. As a result, it is advantageous to operate M_1 and M_2 in the subthreshold regime where a transistor's g_m is maximized for a given current level. Therefore, M_1 and M_2 need to have large W/L ratios. The lengths of M_1 and M_2 then need to be small such that their widths stay relatively small and the input capacitance of the amplifier is not too large.

In order to make sure that all the incremental current caused by the differential input goes through the sources of M_7 and M_8, we cascode the input differential-



(a) Source Admittance (b) Cascode Admittance

Figure 2-6: Circuit schematics for obtaining admittance formula.

pair transistors with M_3 and M_4 to increase their output impedances. The source-degenerated current sources formed by M_5 and R_1 and by M_6 and R_2 are designed to have large output impedances as well. The output impedances of the cascoded input-differential pair and the source-degenerated current sources need to be much larger than the impedance looking into the sources of M_7 and M_8 such that G_m is near g_{m1}.

Before we analyze the operation of the OTA in Fig. 2-3, we shall briefly review two useful admittance formulas. The first one is the formula for the admittance looking into the source of an MOS transistor when its drain is connected to an impedance to incremental ground as shown in Fig. 2-6(a). The second useful formula is the admittance looking into the drain of a cascode transistor as shown in Fig. 2-6(b). Using a nodal analysis, we obtain the two admittances to be

$$G_s = \frac{i_s}{v_s} = \frac{g_{s1} + 1/r_{o1}}{1 + Z_L/r_{o1}}, \tag{2.3}$$

$$G_d = \frac{i_d}{v_d} = \frac{1}{r_{o1}} \cdot \left(\frac{1}{1 + g_{s1}Z_L + Z_L/r_{o1}} \right) \tag{2.4}$$

Let G_{s3} be the admittance looking into the sources of M_3 and M_4, G_{d5} be the admittance looking into the drains of M_5 and M_6, and G_{s7} be the admittance looking into the sources of M_7 and M_8 of the OTA in Fig. 2-3. We can express the transconduc-

tance G_m of the OTA as

$$G_m = g_{m1} \cdot \left(\frac{G_{s7}}{G_{s7} + G_{d5}}\right)\left(\frac{G_{s3} r_{o1}}{1 + G_{s3} r_{o1}}\right). \tag{2.5}$$

We can express G_{s3}, G_{s7} and G_{d5} by using (2.3) and (2.4) as

$$G_{s3} = \frac{g_{s3} + 1/r_{o3}}{1 + 1/\left(r_{o3}(G_{s7} + G_{d5})\right)} \tag{2.6}$$

$$\approx \frac{g_{s3}}{1 + 1/\left(r_{o3}(G_{s7} + G_{d5})\right)}, \tag{2.7}$$

$$G_{s7} = \frac{g_{s7} + 1/r_{o7}}{1 + (1/g_{m11})/r_{o7}} \tag{2.8}$$

$$\approx \left(\frac{g_{m11} r_{o7}}{1 + g_{m11} r_{o7}}\right) \cdot g_{s7}, \tag{2.9}$$

and

$$G_{d5} = \frac{1}{r_{o5}} \frac{1}{1 + R_1/r_{o5} + g_{s5} R_1} \tag{2.10}$$

where g_{si} and r_{oi} are the incremental source admittance of M_i with its drain at incremental ground, and the output resistance of M_i respectively. The expressions from (2.7)-(2.10) present the design constraints for sizing and biasing each device to achieve G_m close to g_{m1}. The size, the channel current and the simulated intrinsic gain ($g_s r_o$) of each transistor in the OTA are shown in Table 3.1. From (2.5), in order to make G_m close to g_{m1}, the ratios $G_{s7}/\left(G_{s7} + G_{d5}\right)$ and $G_{s3} r_{o1}/\left(1 + G_{s3} r_{o1}\right)$ should be made as close to 1 as possible. The ratio $G_{s7}/\left(G_{s7} + G_{d5}\right)$ represents the incremental current gain from the drain of M_3 and M_4 to the output. The incremental current gain from the input differential pair transistors to the drain of the cascode transistors M_3 and M_4 is $G_{s3} r_{o1}/\left(1 + G_{s3} r_{o1}\right)$.

In order to maximize the ratio $G_{s7}/\left(G_{s7} + G_{d5}\right)$, we try to make $G_{d5} \ll G_{s7}$. Since M_{11} and M_7 have the same channel current, $g_{m11} \approx g_{m7}$. Therefore, $g_{m11} r_{o7} \approx g_{m7} r_{o7} \gg 1$ and we have $G_{s7} \approx g_{s7}$. In order to make $G_{d5} \ll G_{s7}$, we need to minimize G_{d5}. From (2.10), we can minimize G_{d5} by making r_{o5} large and also making

$g_{s5}R_1 \gg 1$. Therefore, we make M_5 and M_6 with large W/L ratios and with long channel lengths to achieve large g_{s5} and r_{o5} respectively. Then we choose R_1 such that $g_{s5}R_1 \gg 1$.

In order to maximize the ratio $G_{s3}r_{o1}/(1+G_{s3}r_{o1})$, we need to make $G_{s3}r_{o1} \gg 1$. From (2.7), G_{s3} is approximately g_{s3} if $G_{s7}r_{o3}$ is much greater than 1. Since $G_{s7} \approx g_{s7}$, we have $G_{s7}r_{o3} \approx g_{s7}r_{o3}$. Since the current in M_7 is about 1/16 of the current in M_3 and both transistors are operating in subthreshold, $g_{s7} \approx g_{s3}/16$. From simulation, we achieve $g_{s3}r_{o3}$ of 119 which results in a $g_{s7}r_{o3}$ of 7.43. The expression in (2.7) is thus reduced to $G_{s3} \approx 0.88g_{s3}$. Note that M_1 and M_3 have the same currents and the same channel lengths. Thus M_1 and M_3 should have $r_{o1} = r_{o3}$. As a result, $G_{s3}r_{o1} \approx G_{s3}r_{o3} \approx (0.88g_{s3})r_{o3} = 104$. Therefore, the ratio $G_{s3}r_{o1}/(1+G_{s3}r_{o1})$ is close to 1. As a result, we are able to achieve G_m close to g_{m1} even with sixteen-fold current scaling between the input differential-pair transistors and the folded-branch transistors.

2.2.2 OTA Noise Analysis

The noise in cascode transistors typically contributes little to the overall noise in an OTA [54] because these transistors self shunt their own current noise sources: A cascode transistor's current noise is attenuated by a factor of $1/(1+g_sR)^2$ where g_s is its incremental source transconductance and R is the effective source-degeneration resistance respectively. Therefore, the only noise sources that are significant in Fig. 2-3 are due to non-cascode transistors, i.e., the differential-pair input transistors M_1 and M_2, the resistors R_1 and R_2, and the current-mirror transistors M_{11} and M_{12}. We

Table 2.1: Operating Points for Transistors in the OTA with $I_{tot} = 2.7\ \mu A$

Devices	W/L (μm)	I_D	$g_s r_o$	Operating Region
M_1, M_2	399/1.2	1.18 μA	133	subthreshold
M_3, M_4	100.5/1.2	1.18 μA	119	subthreshold
M_5, M_6	204/6	1.25 μA	322	subthreshold
M_7, M_8	3.6/1.5	68 nA	164	subthreshold
M_9, M_{10}	6/1.2	68 nA	123	subthreshold
M_{11}, M_{12}	3.6/2.2	68 nA	458	above-threshold

now perform an OTA noise analysis using a method similar to that described in [54].

The admittances looking into the sources of M_3, M_5, and M_7 are approximately g_{s3}, g_{s5}, and g_{s7} respectively. Then the current transfer function from each significant current noise source in the OTA to an incrementally grounded output can be calculated to be

$$\frac{\overline{i_{n,out}^2}}{\overline{i_{n,M1}^2}} = \left(\frac{G_{s3}r_{o1}}{1+G_{s3}r_{o1}} \cdot \frac{G_{s7}}{G_{s7}+G_{d5}}\right)^2 \tag{2.11}$$

$$\approx \left(\frac{g_{s3}r_{o1}}{1+g_{s3}r_{o1}} \cdot \frac{g_{s7}}{g_{s7}+G_{d5}}\right)^2, \tag{2.12}$$

$$\frac{\overline{i_{n,out}^2}}{\overline{i_{n,R1}^2}} = \left(\frac{G_{s5}R_1}{1+G_{s5}R_1} \cdot \frac{G_{s7}}{G_{s7}+G_{d3}}\right)^2 \tag{2.13}$$

$$\approx \left(\frac{g_{s5}R_1}{1+g_{s5}R_1} \cdot \frac{g_{s7}}{g_{s7}+G_{d3}}\right)^2, \tag{2.14}$$

and

$$\frac{\overline{i_{n,out}^2}}{\overline{i_{n,M11}^2}} = 1. \tag{2.15}$$

Since this circuit is biased such that $g_{s3}r_{o1} \gg 1$, $g_{s5}R_1 \gg 1$ and $g_{s7} \gg G_{d5}, G_{d3}$ as explained in Section 2.2.1, the expressions from (2.12)-(2.15) are reduced to 1. For the following discussion, we model the MOSFET's current noise as

$$\overline{i_n^2} = 4\gamma kTg_m \tag{2.16}$$

where k is Boltzmann's constant, T is the absolute temperature, g_m is the transconductance of the MOSFET, and $\gamma = 2/3$ for above-threshold operation and $\gamma = 1/(2\kappa)$ for subthreshold operation. From this noise model, we can calculate the input-referred noise of the OTA as the total output current noise divided by its transconductance g_{m1}^2 to be

$$\overline{v_n^2} = \frac{1}{g_{m1}^2}\left(\frac{4kTg_{m1}}{\kappa} + \frac{8kT}{R_1} + \frac{16}{3}kTg_{m11}\right) \tag{2.17}$$

where M_1 and M_2 operate in weak inversion and M_{11} and M_{12} operate in strong

inversion. Let IC be the inversion coefficient of the transistor which is defined as the ratio of its channel current I_D to the moderate inversion characteristic current I_S where I_S is given by [65]

$$I_S = \frac{2\mu C_{ox} U_T^2}{\kappa} \cdot \frac{W}{L} \tag{2.18}$$

where U_T is the thermal voltage and is equal to kT/q, where q is the electron charge. Using the EKV model [16], we can estimate the g_m of each transistor to be

$$g_m = \frac{\kappa I_D}{U_T} \cdot \frac{2}{1 + \sqrt{1 + 4 \cdot \mathrm{IC}}}. \tag{2.19}$$

We can then rewrite (2.17) as

$$\overline{v_n^2} = \frac{1}{g_{m1}} \cdot \frac{4kT}{\kappa} \left(1 + \frac{2U_T}{I_1 R_1} + \frac{4}{3} \kappa\alpha \frac{I_{11}}{I_1} \right) \tag{2.20}$$

where $\alpha = 2/\left(1 + \sqrt{1 + 4 \cdot \mathrm{IC}_{11}}\right)$, which is less than 1, and IC_{11} is the inversion coefficient of M_{11} and M_{12}. Equation (2.20) suggests that in order to minimize the input-referred noise of the OTA, $I_1 R_1$ should be large compared to $2U_T$. Furthermore, the current ratio I_1/I_{11} should be large compared to $\frac{4}{3}\kappa\alpha$. For our implementation, the ratio I_1/I_{11} is 16. For a total supply current of 2.7 μA and 5.3 kHz bandwidth, I_1 and I_{11} are approximately 1.18 μA and 68 nA respectively. For $\mathrm{R_1}$ = 240kΩ, the second and the third terms in (2.20) are 1.8×10^{-1} and 5.4×10^{-2} respectively, assuming a temperature of T=300 K, $\kappa = 0.7$ and $\alpha = 1$. Equivalently, (2.17) is reduced to

$$\overline{v_n^2} = \frac{2kT}{\kappa g_{m1}} \times 2.47. \tag{2.21}$$

Equation (2.21) can be interpreted as 2.47 times the input-referred noise of a MOS transistor operating in weak inversion with a transconductance of g_{m1}. This means that our OTA effectively has only 2.47 subthreshold devices that contribute noise. This value is close to the theoretical limit of 2 noise sources in any OTA that uses two subthreshold MOS differential-pair transistors as an input stage. Effectively, our design has almost eliminated all other sources of noise except for that of M_1 and M_2.

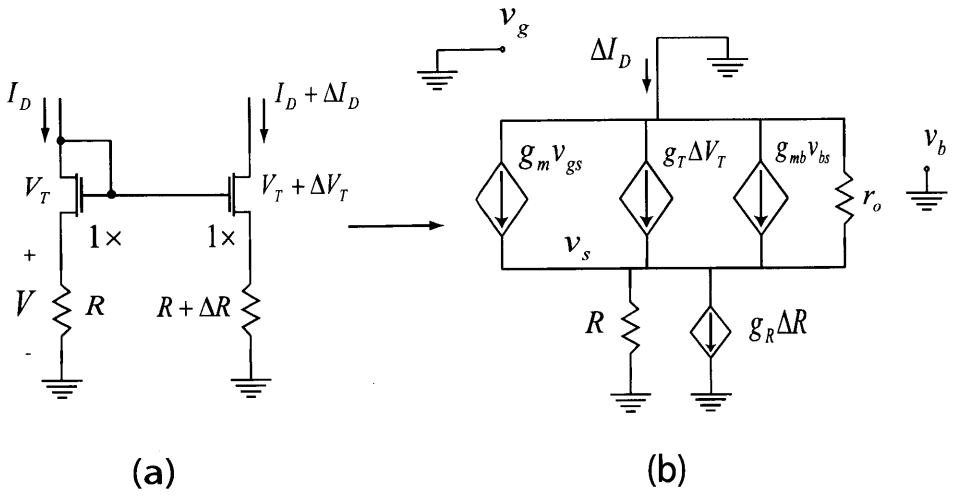


Figure 2-7: Circuit schematics for analyzing V_T and R mismatches in source-degenerated current mirrors.

2.2.3 Current mirror mismatch analysis

The key techniques for achieving good power-noise tradeoff in this amplifier are the uses of source-degenerated current mirrors and the severe current scaling ratio between the input-differential pair transistors and the folded-branch transistors. The severe current scaling scheme can work only if the current errors due to mirroring are well controlled: The amplifier would not work if the error due to current scaling is too large such that none of the current flows in M_7-M_{12} in the OTA of Fig.2-3. Thus, we address and investigate this concern to ensure the correct operation of our amplifier. Let us consider the current matching between two unit transistors in Fig. 2-4 due to variations in the threshold voltage and variations in the source-degeneration resistance. We shall model these variations as errors in the parameters of each of the unit transistors of Fig. 2-4. Let the nominal current in one of the unit transistors of M_{c3} be I_D and consider the deviation in current ΔI_D in one of the unit transistors of M_5 from its nominal value due to deviations in the threshold voltage ΔV_T and deviations in the source-degeneration resistor ΔR as shown in Fig. 2-7(a). To model the threshold-voltage mismatch, we use the body-referenced current equation in saturation for an MOS transistor operating in weak inversion [65]. Let the nominal

current in each unit transistor be described by

$$I_D = I_s e^{\kappa(V_{GS}-V_T)/U_T} \cdot e^{(1-\kappa)V_{BS}/U_T}. \tag{2.22}$$

where I_s is a constant scaling current which is the same for all unit transistors. Let V be the nominal DC voltage drop across R such that $I_D = V/R$. We define

$$g_T = \frac{\partial I_D}{\partial V_T} = -\frac{\kappa}{U_T} \cdot I_D = -g_m, \tag{2.23}$$

$$g_R = \frac{\partial I_D}{\partial R} = \frac{\partial}{\partial R}\left(\frac{V}{R}\right) = -\frac{1}{R} \cdot \frac{V}{R} = -\frac{1}{R} \cdot I_D. \tag{2.24}$$

and

$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = \frac{1-\kappa}{U_T} \cdot I_D. \tag{2.25}$$

By assuming that ΔV_T and ΔR are small, we can use a small-signal circuit model as shown in Fig. 2-7(b) to calculate the variation in nominal current ΔI_D when ΔV_T and ΔR are considered as inputs to the system. With some analysis, the variation in the channel current due to variations in V_T and R is obtained to be

$$\Delta I_D = g_T \cdot \Delta V_T - (g_m + g_{mb} + 1/r_o) \cdot (\Delta I_D - g_R \cdot \Delta R) \cdot R. \tag{2.26}$$

Combining (2.26) with the results from (2.23) and (2.24) and using the relationship $g_s = g_m + g_{mb}$, we obtain the fractional change in channel current as a function of the fractional change in V_T and R to be

$$\frac{\Delta I_D}{I_D} = -\frac{1}{1 + g_s R + R/r_o} \cdot \frac{\Delta V_T}{I_D/g_m} - \frac{g_s R + R/r_o}{1 + g_s R + R/r_o} \cdot \frac{\Delta R}{R}. \tag{2.27}$$

Since M_{c3}, M_5 and M_6 are biased in weak-inversion regime, their I_D/g_m is approximately 40 mV at room temperature. As seen from (2.27), the mismatch in threshold voltage as a fraction of 40 mV is attenuated by a factor of $1 + g_s R + R/r_o$ and is negligible if $g_s R \gg 1$. In our design, we have $g_s R \approx 12$, thus, the fractional mismatch in threshold voltage is attenuated by more than a factor of 10 and does not play

a significant role in current mirror mismatch. In contrast, the fractional mismatch in channel current scales almost 1:1 to the fractional mismatch in R. However, the matching of passive components in most CMOS processes is much better controlled than the matching of transistors' threshold voltages. In our design, therefore, we try to achieve good resistor matching with careful layout.

2.2.4 Noise Efficiency Factor and its theoretical limit for OTA with differential inputs

To compare the power-noise tradeoff among amplifiers, we adopt the noise efficiency factor (NEF) proposed in [63] and widely used to compare neural-amplifier designs:

$$\mathrm{NEF} = V_{ni,rms}\sqrt{\frac{2I_{tot}}{\pi \cdot U_T \cdot 4kT \cdot BW}} \tag{2.28}$$

where $V_{ni,rms}$ is the total input-referred noise, I_{tot} is the total supply current, and BW is the -3 dB bandwidth of the amplifier respectively. The theoretical limit of the NEF of an OTA that uses a differential pair as an input stage is when the two differential-pair transistors are the only noise sources in the circuit. The input-referred noise of the OTA is then $\overline{V_{ni}^2} = 2 \times 2kT/(\kappa g_m) = 4kT/(\kappa g_m)$ where g_m is the transconductance of a single differential-pair transistor. For minimum input-referred noise, the transistors should run in subthreshold, such that we have $g_m = \kappa I_D/U_T$. Assuming a first-order roll-off of the frequency response, the input-referred noise of the ideal OTA is expressed as

$$V_{ni,rms} = \sqrt{\frac{4kT \cdot U_T}{\kappa^2 I_D} \cdot \frac{\pi}{2} \cdot BW}. \tag{2.29}$$

Combining (2.28) and (2.29) and setting $I_{tot} = 2I_D$, we obtain the theoretical limit for NEF of any OTA that uses a subthreshold MOS differential pair to be

$$\mathrm{NEF} = \frac{\sqrt{2}}{\kappa} \approx 2.02 \tag{2.30}$$

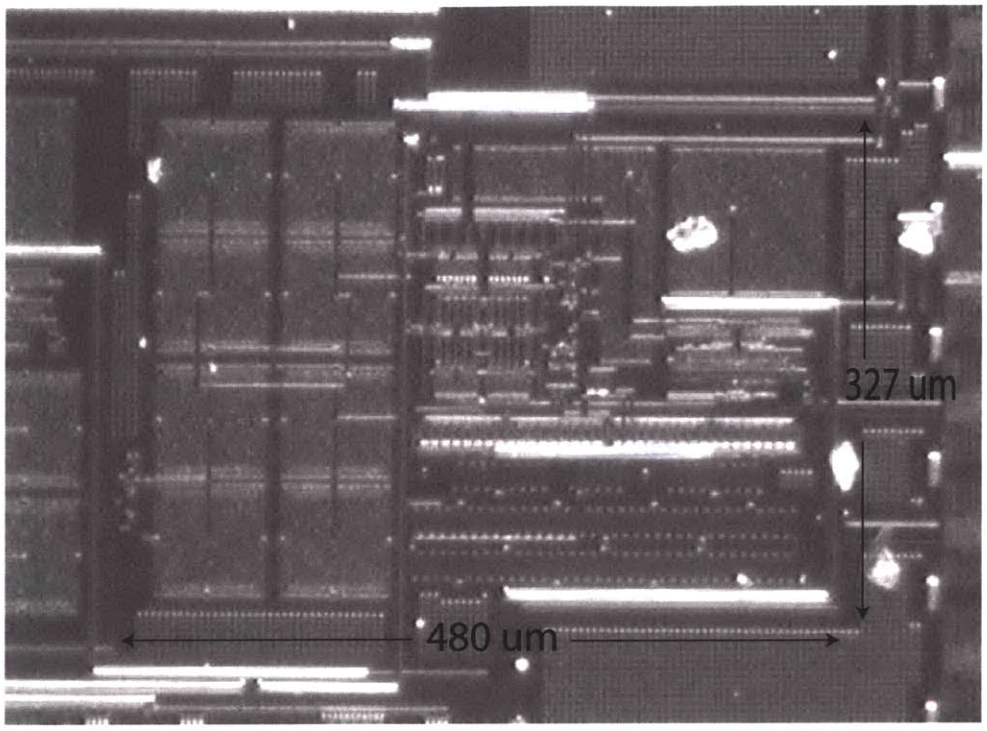


Figure 2-8: A die micrograph of our neural amplifier.

assuming a typical value of $\kappa = 0.7$. We now show that our experimental NEF is near this value, and our theoretical NEF was computed to be 2.47 from Section 2.2.2.

2.3 Measurement Results

The amplifier was fabricated in a 0.5 μm CMOS process through MOSIS. It was designed to give a gain of approximately 110 (40.8 dB) by setting the value of C_{in} to 14 pF and C_f to 120 fF. The OTA in the bandpass filter stage is a wide common-mode range OTA to reduce signal distortion in the case of large input amplitudes. The amplifier occupies a chip area of 0.16 mm^2. A chip micrograph of our amplifier is shown in Fig. 2-8.

Four chips were tested on the lab bench and they exhibited very similar performance characteristics, indicating that the severe current-scaling scheme worked robustly. The measured transfer function of one of our neural amplifiers is shown in Fig. 2-9. The amplifier consumes 2.7 μA including the current from the bias circuit

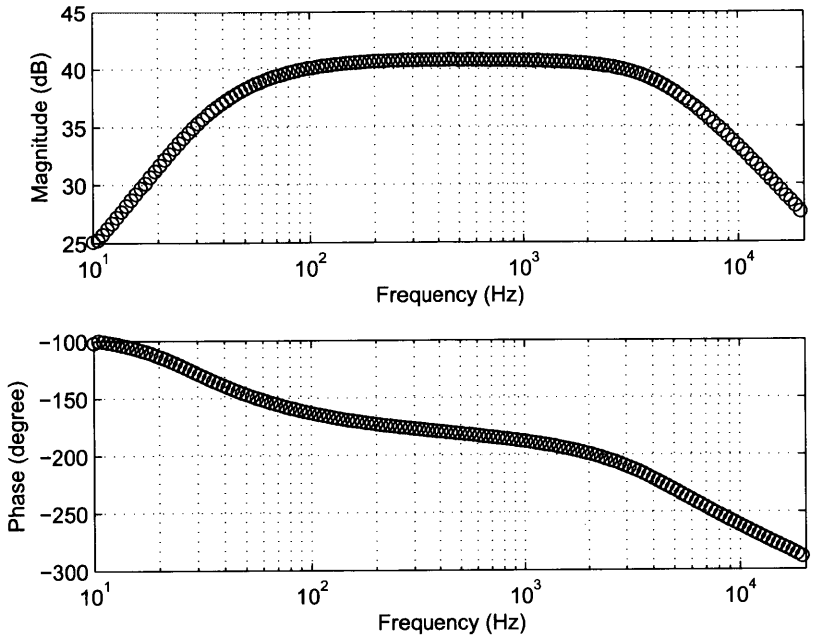


Figure 2-9: Measured transfer function of the neural amplifier configured for recording neural spikes.

(M_{b2}, M_{c2} and M_{c3}) from a 2.8 V supply. We do not include the current I_{bias} shown in Fig. 2-3 since it can be shared by many amplifiers in the array. The -3 dB cutoff frequencies are adjusted to be at 45 Hz and 5.32 kHz. The amplifier is configured as an inverting amplifier, thus the phase is approximately -180° near the midband frequency.

Fig. 2-10 shows the measured input-referred noise spectrum together with a circuit simulation of the noise spectrum with a similar noise model to the theoretical calculations (the smooth curve). There is a good agreement between the measured and simulated curves. The measured thermal noise level is 31 $\text{nV}/\sqrt{\text{Hz}}$. Integrating under the area of the measured curve from 10 Hz to 98 kHz yields a total input-referred noise of 3.06 μV_{rms}, while the simulated result is 3.1 μV_{rms}. With a high-pass cutoff frequency at 45 Hz, 1/f noise is filtered out and is not noticeable in the passband.

The NEF of this amplifier is calculated from the achieved experimental measurements to be 2.67. This value is close to 2.02 which is the theoretical NEF limit that has been calculated in 2.2.4 and also near our expected theoretical calculation of 2.47

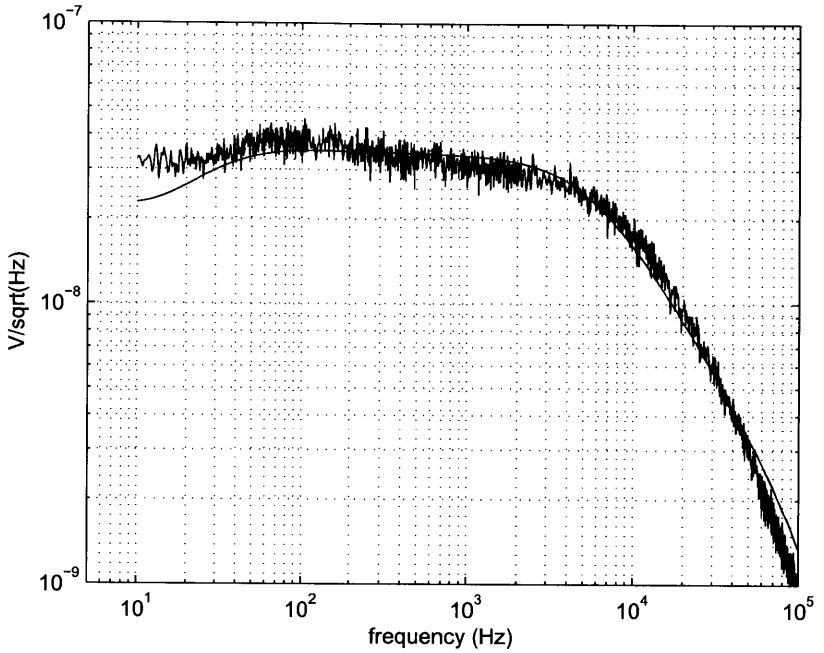


Figure 2-10: Measured and simulated (smooth curve) input-referred noise spectra of the neural amplifier configured for recording neural spikes.

Table 2.2: Comparison of reported neural amplifiers

Amplifier	**Power** μW	**Noise** μV_{rms}	**Bandwidth**	**NEF**	**Year Published**
[24]	80	2.2	0.025 Hz-7.5 kHz	4	2003
[20]	8.6	5.6	100 Hz-9.2 kHz	4.9	2009
[38]	40.3	1.94	0.2 Hz-8.2 kHz	2.9	2009
[72]	0.44-0.9	2.5	3.5 mHz-292 Hz	3.26	2009
[32]	14.8	4.3	10 Hz-9 kHz	5.56	2010
This work	7.56	3.06	45 Hz-5.32 kHz	2.67	2007

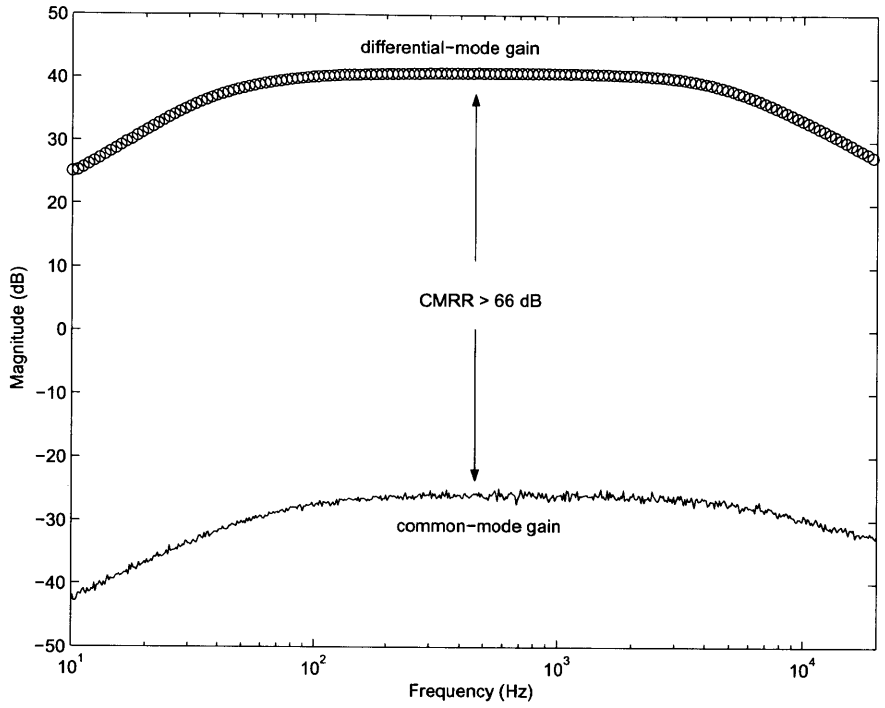
in (2.21). The good power-noise tradeoff of this amplifier is a result of minimizing the effective number of transistors that contribute noise. Moreover, almost all the power is consumed by the input-differential pair. Therefore, little power is wasted in less critical parts of the amplifier. Table 2.2 compares the NEF of this amplifier with those reported in the literature. Due to its energy-efficiency that is close to a limit determined by physics, this amplifier still achieves the best NEF among the more recently reported design, even though it was published earlier.

The measured CMRR and PSRR are shown in Fig. 2-11. CMRR is calculated as the ratio of the differential-mode gain to the common-mode gain. PSRR is calculated as the ratio of the differential-mode gain to the gain from power supply to the output. The measured CMRR and PSRR exceed 66 dB and 75 dB (over the range of 45 Hz to 5.32 kHz) respectively. The measured characteristics of the neural amplifier are summarized in Table 2.3.

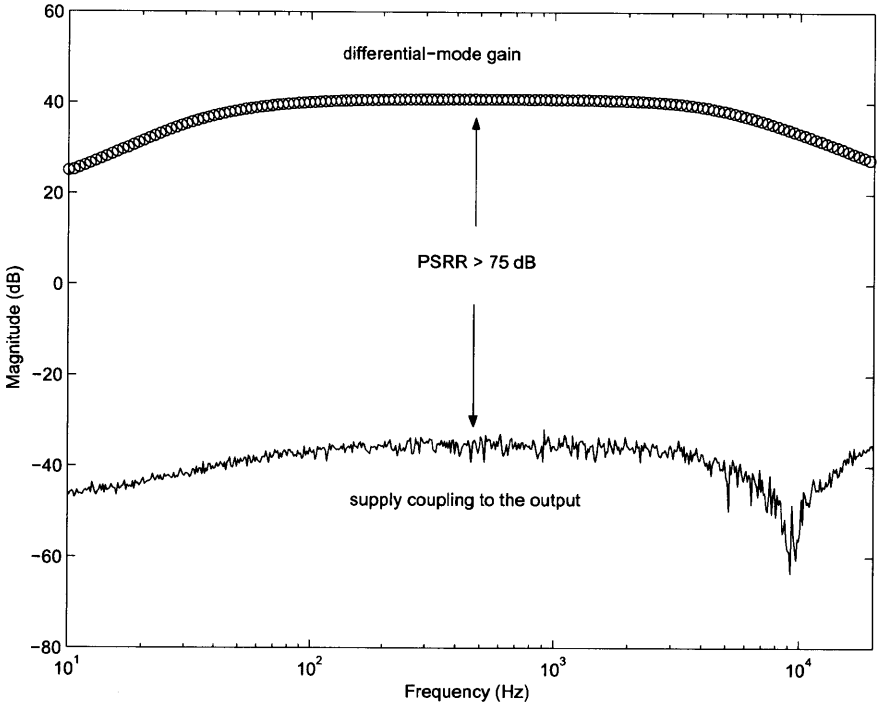
Table 2.3: Measured Performance Characteristics

Parameter	**Measured**
Supply voltage	2.8 V
Total current	2.7 μA
Gain	40.85 dB
Bandwidth	45 Hz-5.32 kHz
Input-referred noise	3.06 μV_{rms}
Noise efficiency factor	2.67
Max. signal (1% THD @ 1.024 kHz)	7.3 mV_{pp}
Dynamic Range (1% THD)	58 dB
CMRR (45 Hz-5.32 kHz)	66 dB
PSRR (45 Hz-5.32 kHz)	75 dB
Area (in 0.5 μm CMOS)	0.16 mm^2

We verified that this neural amplifier works in a real recording environment by using it to record action potentials in the RA region of a zebra finch's brain. Data were taken with a Carbostar electrode that had an impedance of approximately 800 kΩ. A long extracellular trace and a short extracellular trace recorded from our amplifier normalized by the gain are shown in Fig. 3-49. They were found to be identical to that recorded by a commercial neural amplifier.

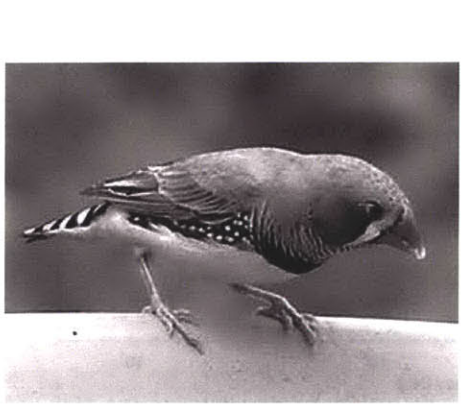


(a) CMRR

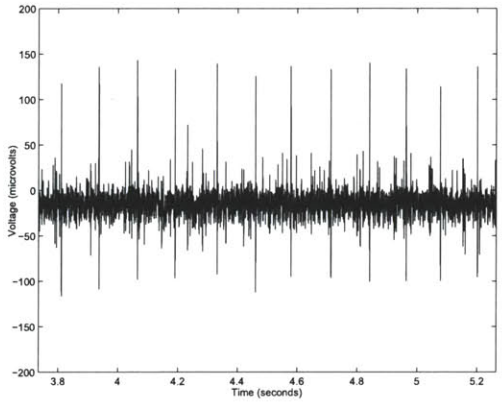


(b) PSRR

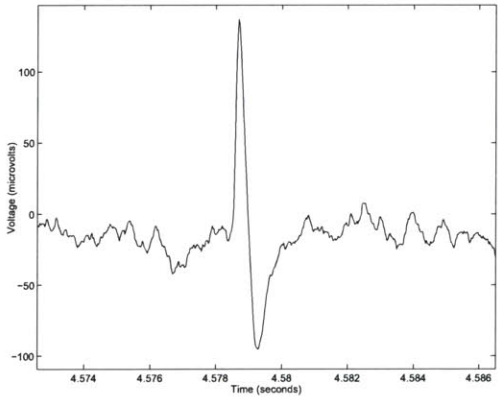
Figure 2-11: CMRR and PSRR measurements of the neural amplifier configured for recording action potentials.



(a) A zebra finch



(b) Spike train



(c) Single spike

Figure 2-12: Neural recording from a zebra finch's brain: (a) A zebra finch (b) Long time trace (c) Short time trace.

2.4 Measurements of Local Field Potentials

Local Field Potentials (LFPs) instead of action potentials are often used in some brain-machine interfaces, e.g, those used in paralysis prosthetics [56]. Therefore, we also measured the performance characteristics of our amplifier configured with lower bandwidth (and power) for such applications. Since the LFP contains energy in the frequency range of 1 Hz to 300 Hz, we can simply lower the -3 dB lowpass cutoff frequency of our amplifier by lowering the supply current of the OTA in the bandpass filter stage. The highpass cutoff frequency can also be lowered to be below 1 Hz by adjusting V_{tune}. If we just change the bandwidth in this manner, the input-referred noise of the amplifier becomes excessively low. From a hand-analysis, if we adjust the bandwidth of the amplifier to be 0.5 Hz-300 Hz while maintaining the same supply current for the gain-stage OTA, the input-referred noise of the amplifier is less than 1 μV_{rms}. Such low input-referred noise is unnecessary and is wasteful of power. From (2.20), the input-referred noise power is inversely proportional to g_{m1}, therefore inversely proportional to I_1. Thus, we can save more power by lowering the current in the gain-stage OTA as well.

The amplifier was adjusted to have a highpass cutoff frequency of 392 mHz and a lowpass cutoff frequency of 295 Hz for LFP-suitable configuration. The total current of our amplifier was measured to be 743 nA, corresponding to a power consumption of 2.08 μW from a 2.8 V supply and 1.66 μV_{rms} total input-referred noise integrated from 0.2 Hz to 1 kHz. The measured transfer function for the amplifier configured for recording LFP is shown in Fig. 2-13. The measured input-referred noise spectrum and expected noise curve from simulation are shown in Fig. 2-14. The measured NEF for LFP recording is then 3.21. Note that the NEF is worse than that of the amplifier configured to record neural spikes. This degradation in NEF is due to the fact that the thermal noise from the resistors R_1 and R_2 becomes more significant once the current in the input differential pair is low. Moreover, 1/f noise becomes significant as well since the highpass cutoff has been decreased to 395 mHz. The other measured performance characteristics of the LFP amplifier are summarized in Table 2.4 and

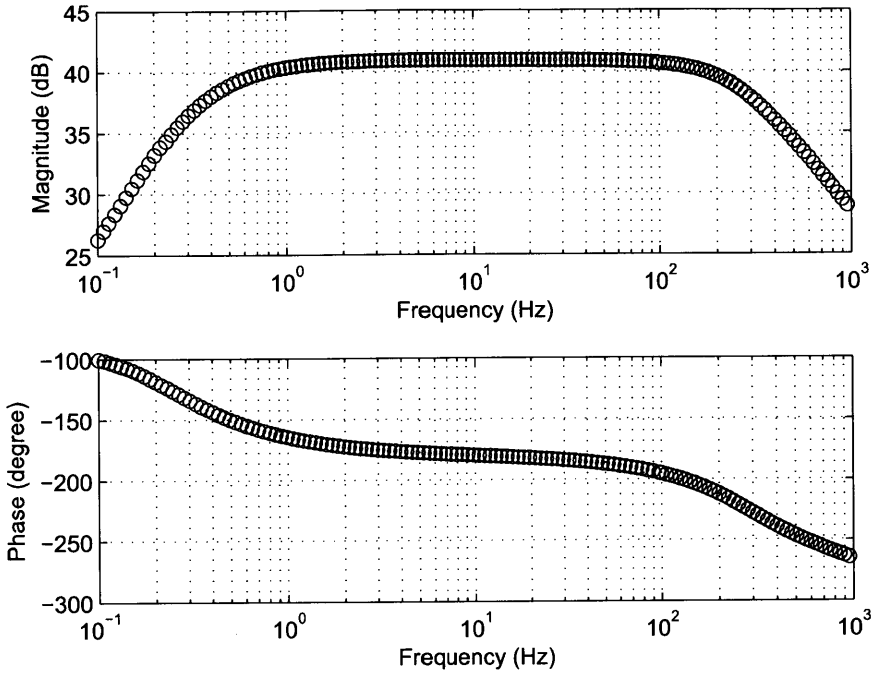


Figure 2-13: Transfer function of the amplifier configured for recording LFP.

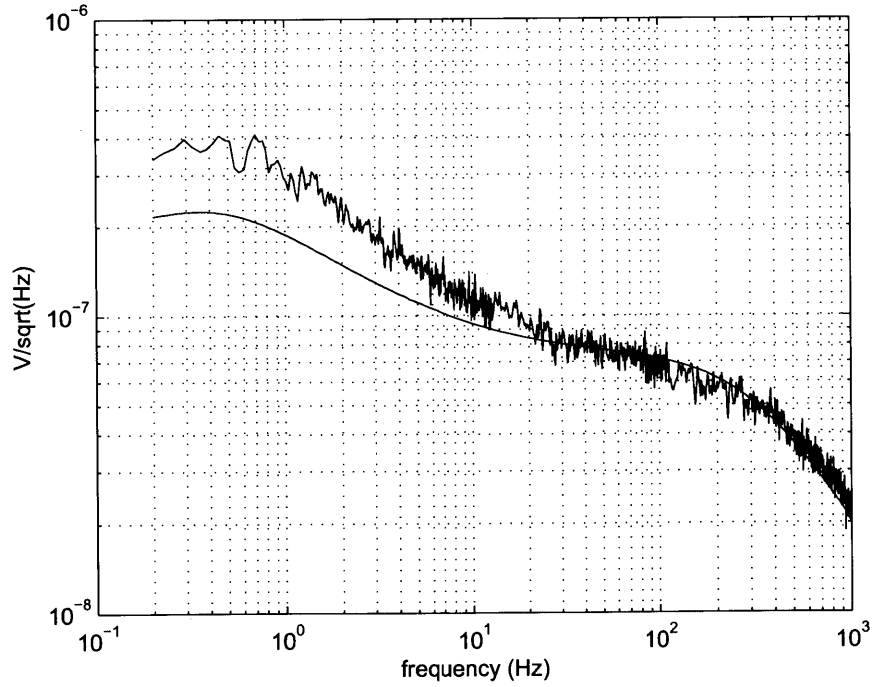


Figure 2-14: Measured and simulated (smooth curve) input-referred noise spectra for the amplifier configured for recording LFP.

similar to those shown in Table 2.3.

Table 2.4: Measured Performance Characteristics of LFP Amplifier

Parameter	Measured
Supply voltage	2.8 V
Total current	743 nA
Gain	40.9 dB
Bandwidth	392 mHz-295 Hz
Input-referred noise	1.66 μV_{rms}
Noise efficiency factor	3.21
Max. signal (1% THD @ 1.024 kHz)	7.2 mV_{pp}
Dynamic Range (1% THD)	63.7 dB
CMRR (392 Hz-295 Hz)	66 dB
PSRR (392 Hz-295 kHz)	75 dB
Area (in 0.5 μm CMOS)	0.16 mm^2

2.5 Conclusion

This chapter presented a micropower low-noise neural recording amplifier. Many low-noise design techniques were employed to enable the amplifier to achieve an input-referred noise near the theoretical limit of two devices of an input differential pair. The amplifier appears to be the lowest power and most energy-efficient neural amplifier reported to date. It can be configured to record either action potentials or local field potentials. We obtained successful recordings of action potentials with our amplifier from a zebra finch's brain. This amplifier may thus be useful in brain-machine interfaces for paralysis prosthetics, visual prosthetics, or experimental neuroscience systems for chronic monitoring.

Chapter 3

Ultra-low-power 32-channel Neural Recording IC

To record from a large number of cortical neurons, high-channel-count recording systems are needed. In such case, low power consumption and small area per recording channel are of critical importance. In this chapter, we present a design and experimental results of an ultra-low-power 32-channel neural recording IC. The 32-channel neural recording IC is the front-end processing core of the internal unit. Its function is to amplify and digitize neural signals from 32 recording electrodes, and send the digitized neural data to the FPGA on the internal unit for further processing before the processed neural data is transmitted to the external unit via the data telemetry system. The organization of this chapter is as follows. In Section 3.1, we give an overview of the overall system architecture of the neural recording IC. In Section 3.2, we discuss the design of the neural amplifier. In Section 3.3, we present detailed designs of an energy-efficient analog-to-digital converter (ADC), an analog multiplexer, and the power saving strategies that are applied to these circuit building blocks. In Section 3.5, we discuss the serial programming protocol for configuring the IC, and also the circuit architecture of the serial programming interface unit. In Section 3.6, we present the design of the Digital Control Unit that oversees the operation of the whole chip. Finally, in Section 3.7, we present both the measurement results of the circuit building blocks, and also the *in-vivo* experimental result of the 32-channel

neural recording IC, when used in a wireless neural recording experiment to obtain neural signals from a behaving primate.

3.1 System Architecture

Figure 3-1 shows the overall architecture of the 32-channel neural recording IC. The IC contains a total of 32 recording channels, which are grouped into eight 4-channel neural recording modules. The schematic of one of the 4-channel neural recording modules is shown in Fig 3-2. Each neural recording module contains four neural amplifiers, an analog multiplexer, an 8-bit ADC, and a serial programming interface unit. The outputs from the four neural amplifiers in the neural recording module are multiplexed into the ADC which digitizes its input signal at a rate of 125 kS/s. Effectively, each neural amplifier's output is sampled and digitized at a rate of 31.25 kS/s. The clock and control signals for the analog multiplexer and the ADC are generated from a centralized control logic which we call the Digital Control Unit. The output data from the ADCs are multiplexed into the Digital Control Unit, where the data are packetized and streamed off-chip for further processing by the on-board FPGA. The configuration setting of each recording channel is achieved through the serial programming interface unit via the programming data and the programming clock pins.

To minimize power consumption of the recording system, we utilize two power supply domains. The neural amplifiers and analog multiplexer, requiring larger voltage headroom, operate from a 1.8 V supply voltage. The ADCs and the Digital Control Unit operate from a lower supply voltage of 1 V to save power. The digital level translators are included to interface between the Digital Control Unit and the control switches in the analog multiplexers. The analog multiplexers also act as DC level shifters between the neural amplifiers and the ADCs. A bandgap voltage reference circuit [31, 68] is included on chip to generate a temperature-independent 1 V reference (V_{ref}) for the ADCs and a 0.9 V reference (V_{mid}) for the mid-rail voltage of the neural amplifiers. A proportional-to-absolute-temperature (PTAT) current gen-

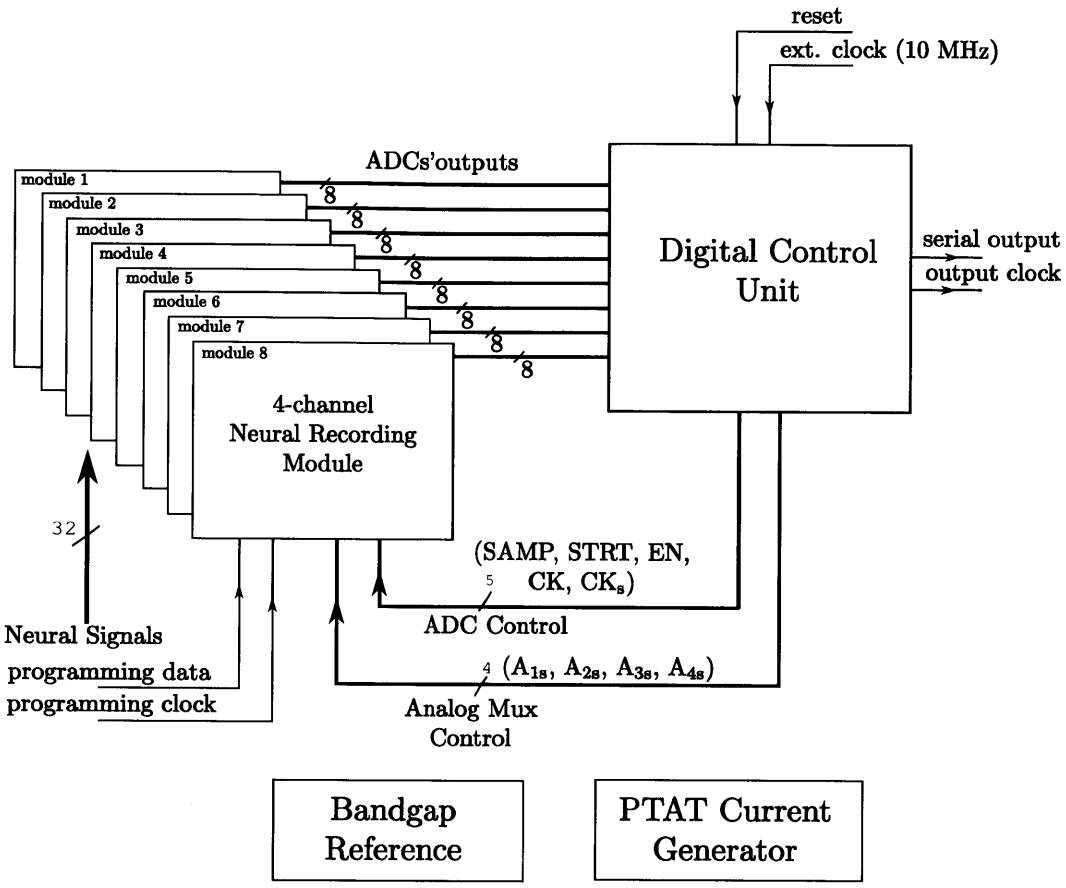


Figure 3-1: Overall architecture of the 32-channel neural recording system.

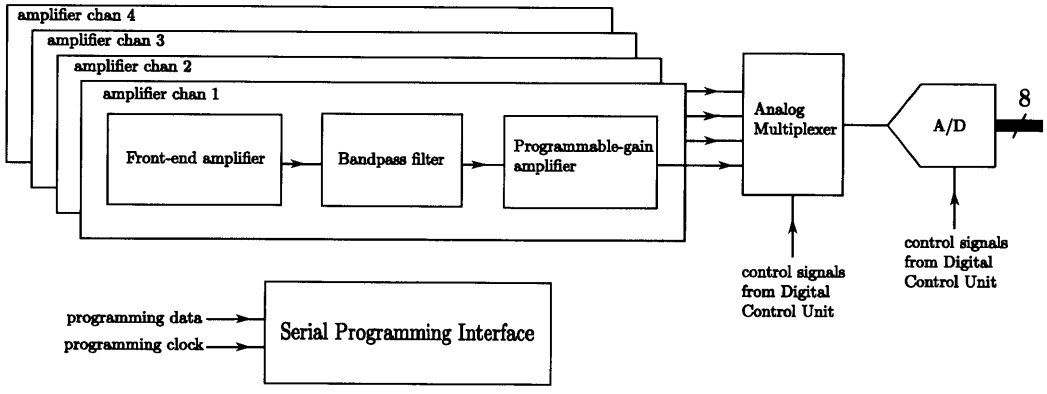


Figure 3-2: Architecture of a 4-channel neural recording module.

erator [21] provides constant-gm biasing to all the neural amplifiers and the analog multiplexers on the IC.

3.2 Neural Amplifier

While the amplifier presented in Chapter 2 consumes very low power and exhibits very low input-referred noise, its layout area of 0.16 mm^2 might prevent it from a practical use in a high-channel-count neural recording system. For such system, low power consumption and small area per channel are of critical importance. In this section, we describe the design of the neural amplifier used in the 32-channel neural recording IC. The amplifier is modified from the one presented in Chapter 2 to occupy smaller silicon area and to provide programmable gain and bandwidth. The neural amplifier is optimized for recording action potentials (spikes), which are widely considered to be the main information carrying signals in the brain. However, local field potentials (LFPs) have been shown to provide promising additional neural information for the development of BMI systems [6, 56]. It is therefore beneficial if neural amplifiers in the array can also be used to record the LFPs. Therefore, we designed our neural amplifiers such that they can also be configured to record LFPs if needed. Generally, extracellular spikes exhibit frequency content from 200 Hz - 10 kHz with amplitudes in the range of 10 μV - 500 μV, while LFPs exhibit frequency content in the range of $<$ 1 Hz-300 Hz with amplitudes in a few millivolts range ($<$ 5 mV). With the neural noise due to background cortical activity on the order of 5-10 μV_{rms} [22], the SNR of the combined neural signals is on the order of 60 dB. To record both LFPs and spikes, a neural amplifier should provide an effective input dynamic range of 60 dB, which is determined by the largest neural signal amplitude (LFPs) to that of the lowest signal amplitude (low-SNR spikes). To achieve high input dynamic range while keeping the supply voltage relatively low, the neural amplifier's gain should be made programmable. Spiking signals and LFPs can be separated in the frequency domain, allowing each signal type to be amplified appropriately with different gains [45]. Weaker spikes should be amplified with high gains such that

noise from subsequent signal processing stages do not degrade the relatively lower-SNR spike signals. For LFPs with relatively higher SNR, lower gain can be used such that the LFPs do not saturate the output of the neural amplifier.

To get clean neural recordings, the input-referred noise of the amplifier should be kept low. A common design choice for most existing neural amplifiers in the literature is that their input-referred noise be kept below the background noise that may be encountered at any recording electrode in the array. In real recording environments, the background noise strength encountered at various electrodes in the array may vary considerably. Figure 3-3 shows an example of a probability distribution of the background noise obtained from an array of 64 electrodes. This distribution shows that while some recording sites can exhibit a background noise as low as 5 μV_{rms}, most recording sites exhibit a background noise higher than 15 μV_{rms}. In a thermal-noise-limited subthreshold amplifier, amplifier's power consumption for a fixed-bandwidth signal scales as $1/v_n^2$ where v_n is the input-referred noise of the amplifier. Such a power-noise tradeoff shows a steep power cost of achieving very low input-referred noise in an amplifier. Therefore, biasing every amplifier in an array such that its input-referred noise is below 5 μV_{rms}, the lowest background noise encountered over all sites, is wasteful of power. In order to optimize power consumption in the neural recording system, an adaptive biasing scheme should be used [55]. In such a scheme, each neural amplifier's input-referred noise, and thus its power consumption, can be individually adjusted to suit the background noise level at its corresponding recording site. As a result, every neural amplifier in the array consumes just sufficient power to obtain clean recordings, while the total power consumption of the recording system is near optimal.

Figure 3-4 shows the schematic of the neural amplifier. The neural amplifier consists of three stages including: i) the front-end amplifier ii) the bandpass filter and iii) the programmable-gain amplifier. The midband gain of the front-end amplifier is designed to be 40 dB. The passband of the neural amplifier is determined by that of the bandpass filter stage and can be chosen for one of the following two settings: i) the spike-recording setting (350 Hz - 12 kHz) and ii) the LFP recording setting ($<$ 1 Hz

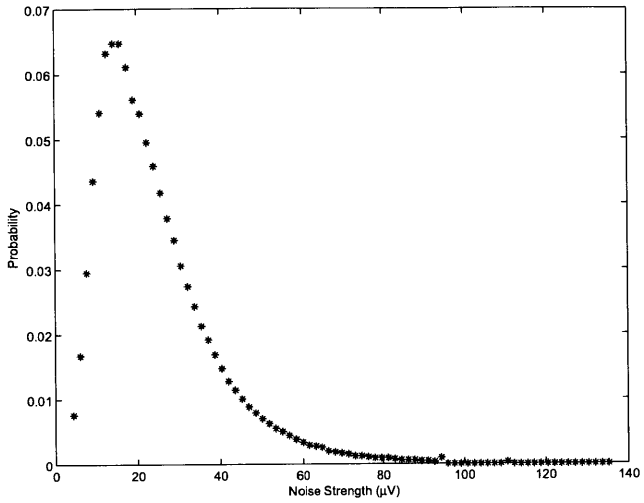


Figure 3-3: An example of a probability distribution of the neural background noise measured from a recording array of 64 electrodes.

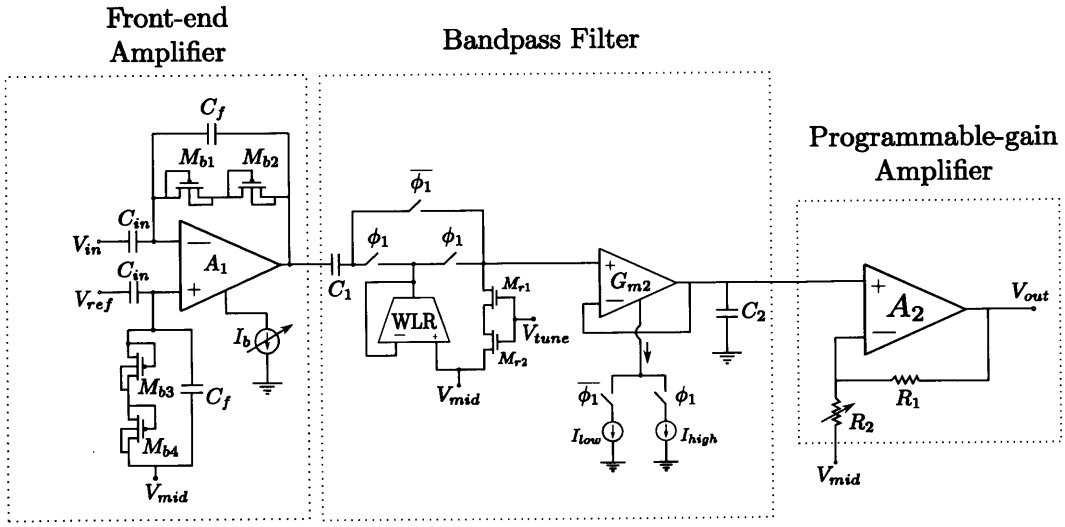


Figure 3-4: Schematic of the neural amplifier consisting of three stages: i) front-end amplifier ii) bandpass filter iii) programmable-gain amplifier.

- 300 Hz). After the bandpass filter stage, the programmable gain amplifier provides an additional gain that ranges from 9 dB to 26 dB, adjustable in eight unequal steps. As a result, the overall gain of the neural amplifier can be adjusted from 49 dB to 66 dB based on a user-provided digital input.

3.2.1 Front-End Amplifier

Since the front-end amplifier is the first stage of the whole signal processing chain, its input-referred noise is of critical importance. Figure 3-5(a) shows the schematic of the front-end amplifier. We use the capacitively-coupled architecture proposed in [24] to reject the DC offset introduced at the electrode-tissue interface. The midband gain of the front-end amplifier is set to $-C_{in}/C_f$ which is achieved by the capacitive feedback formed by C_f and C_{in} around the high gain amplifier A_1. The high-resistance pseudoresistor element formed by M_{b3} and M_{b4} acts as a DC biasing resistor to set the DC voltage at the positive terminal of A_1 to the amplifier's mid-rail voltage V_{mid}. The pseudoresistor element formed by M_{b1} and M_{b2} provides a DC feedback path from the output of the front-end amplifier, V_{out}, to the negative input terminal of A_1 such that the voltages at V_{out} and the negative input terminal of A_1 are set to V_{mid} in steady state.

The schematic of the amplifier A_1 including its transistor sizing is shown in Figure 3-5(b). The amplifier A_1 consists of a folded cascode operational transconductance amplifier (OTA) followed by a class-AB output buffer. The class-AB output buffer is included to minimize the output impedance of A_1 while minimizing the extra quiescent bias current required to operate the buffer. This feature is important to ensure that even at a low bias current level of the OTA, the closed-loop bandwidth of the front-end amplifier is still much wider than the bandwidth of the bandpass filter. Therefore, the bandwidth of the overall neural amplifier is determined by that of the bandpass filter stage, and is not limited by the bandwidth of the front-end amplifier even if its bias current is drastically reduced to save power according to the adaptive biasing strategy. The transistors M_{b1}, M_{b2}, M_1 - M_{14} form the core of the OTA while M_{15}-M_{18} form the class-AB output buffer. The transistors M_{b3}-M_{b8} form the bias cir-

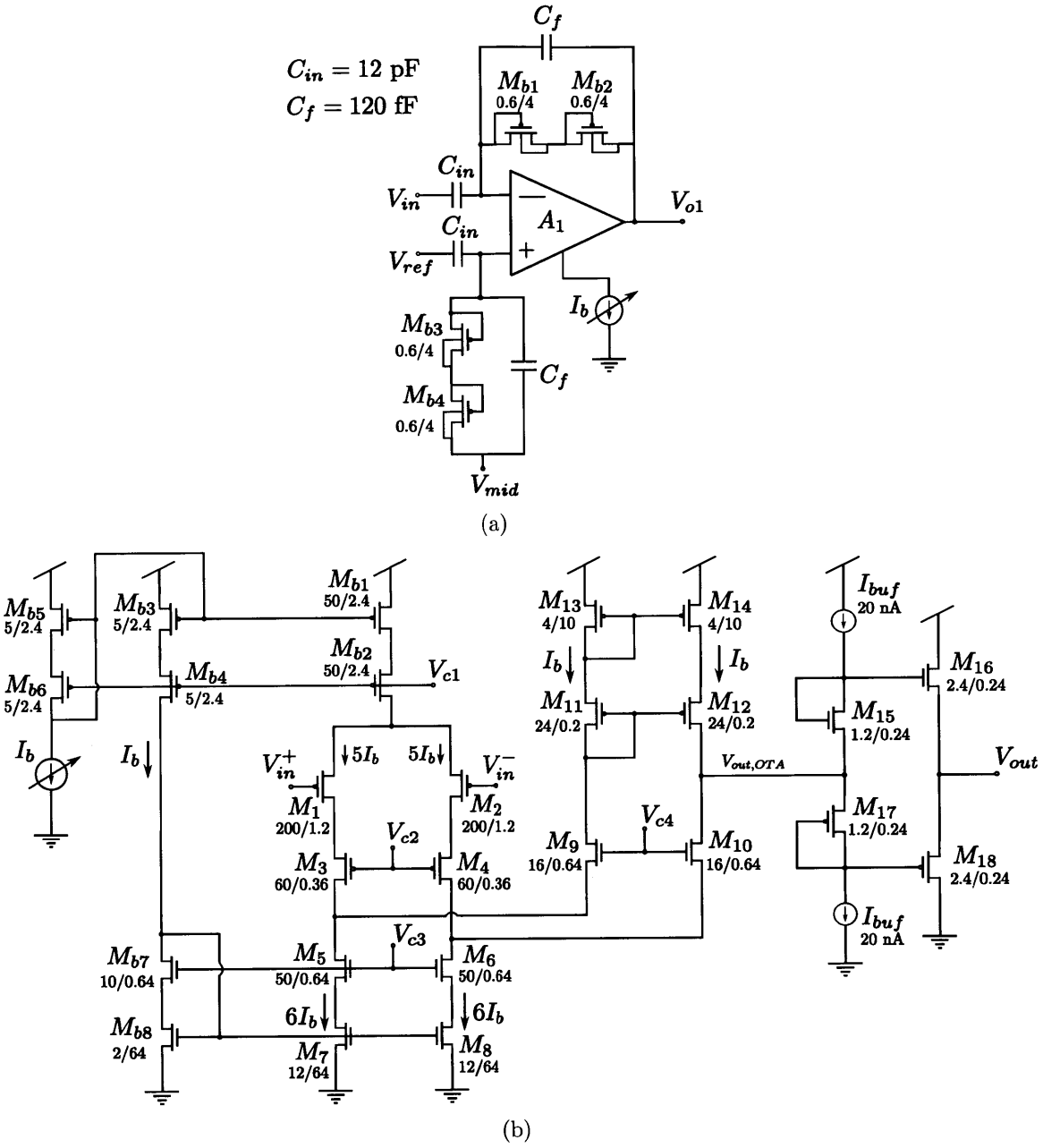


Figure 3-5: Schematic of the front-end amplifier: (a) High-level schematic (b) Schematic of the amplifier A_1.

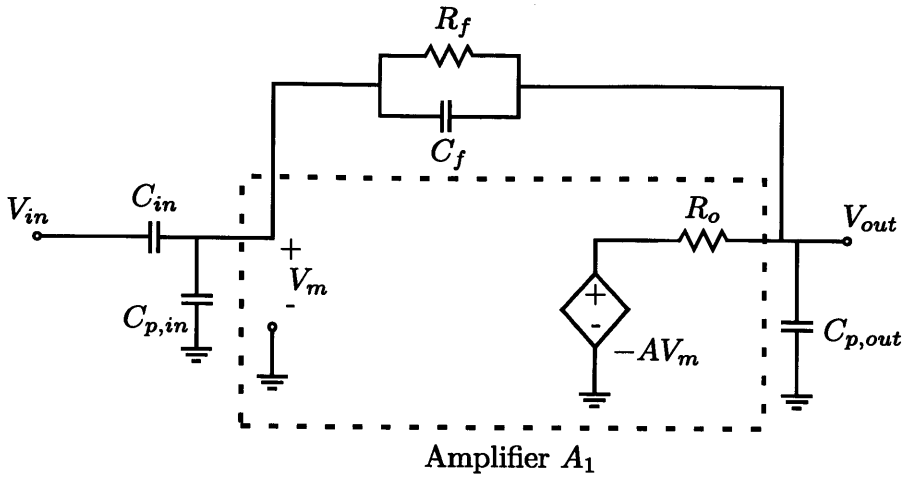


Figure 3-6: Small-signal diagram of the front-end amplifier.

cuit that helps distribute the current in the folded-cascode OTA. The distribution of current in the OTA will be explained later in this section. The folded-cascode OTA is modified from the low-power, low-noise OTA presented in Chapter 2 (and in [69]). To minimize the layout area per recording channel, we replaced the source-degeneration resistors in the OTA in [69] with the transistors M_{b8}, M_7, and M_8 at an expense of reduced noise efficiency factor. The cascode voltages V_{c1}, V_{c2}, V_{c3} and V_{c4} and the current I_{buf} for biasing the output buffer are generated from bias circuits local to the OTA (not shown in the figure for simplicity).

Small-Signal Analysis

In this section, we will analyze the small-signal operation of the front-end amplifier. The analysis provided here is important to determine the input-referred noise and the bandwidth of the overall neural amplifier. The small-signal diagram of the front-end amplifier is shown in Fig. 3-6. We will use the block diagram technique as explained in [51] since it provides the most intuitive way of analyzing a feedback circuit. The resistance R_f represents the effective resistance of the series combination of M_{b1} and M_{b2} in Fig. 3-5(a). We can model the amplifier A_1 as an amplifier with an infinite input resistance, a transfer function of $A(s)$, and an output resistance of R_o. The capacitive load at the output node V_{out} and the parasitic capacitance at the input

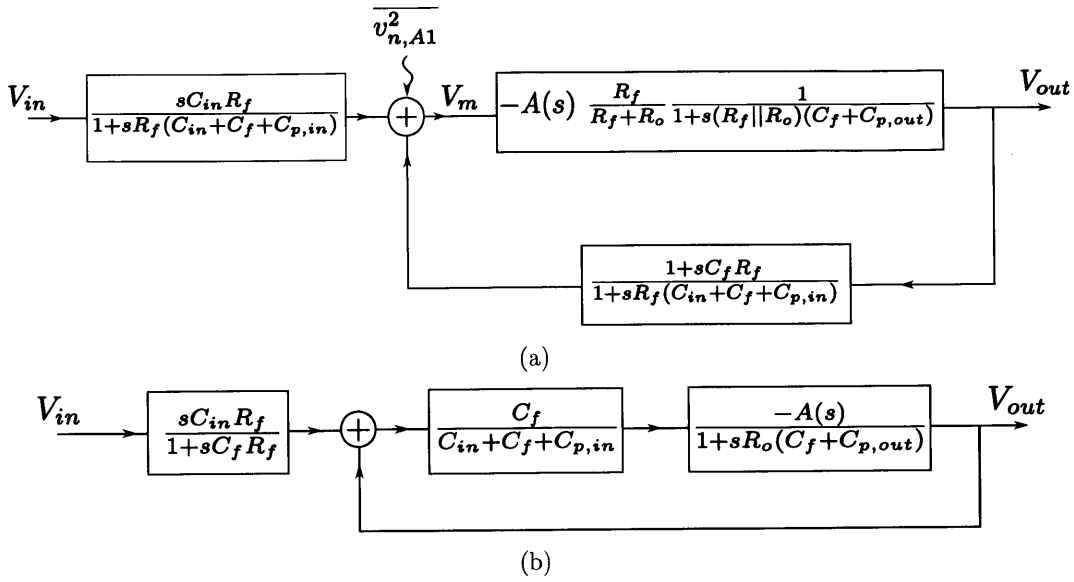


Figure 3-7: Feedback block diagram for analyzing the front-end amplifier.

terminal of A_1 can be modeled as $C_{p,out}$ and $C_{p,in}$ respectively. The incremental voltage across the input terminals of A_1 is $V_{in}^- - V_{in}^+ = V_m$. Using the principle of superposition, we can write the expression for V_m as

$$\begin{aligned} V_m &= V_{in}\frac{R_f||\,(1/s(C_f+C_{p,in}))}{1/sC_{in}+R_f||\,(1/s(C_f+C_{p,in}))} + V_{out}\frac{1/s(C_{in}+C_{p,in})}{1/s(C_{in}+C_{p,in})+R_f||\,(1/sC_f)} \\ &= V_{in}\frac{sC_{in}R_f}{1+sR_f(C_{in}+C_f+C_{p,in})} + V_{out}\frac{1+sC_fR_f}{1+sR_f(C_{in}+C_f+C_{p,in})}. \end{aligned} \tag{3.1}$$

The expression for V_{out} can be approximated as

$$\begin{aligned} V_{out} &= -A(s)V_m\frac{R_f||\,(1/s(C_{p,out}+C_f))}{R_o+R_f||\,(1/s(C_{p,out}+C_f))} \\ &= -A(s)V_m\frac{R_f}{R_f+R_o}\frac{1}{1+s(R_f||R_o)(C_f+C_{p,out})}. \end{aligned} \tag{3.2}$$

Using (3.1) and (3.2), we can draw a feedback block diagram for the front-end amplifier as shown in Fig. 3-7(a). In Fig. 3-7(a), we have included the input-referred noise of the amplifier A_1, $\overline{v_{n,A1}^2}$, for the purpose of noise analysis which will be presented later. Since R_f is the effective resistance of the pseudoresistor element, its value can be much larger than R_o (R_f is on the order of 10^{12} ohms [24]). Thus,

$R_f/(R_f + R_o) \approx 1$ and $R_f||R_o \approx R_o$. Therefore, the expression in (3.2) can be simplified to

$$V_{out} = -A(s)V_m \frac{1}{1 + sR_o(C_f + C_{p,out})}. \tag{3.3}$$

Figure 3-7(a) can be simplified into a unity-feedback configuration as shown in Fig. 3-7(b). To arrive at Fig. 3-7(b), we have made approximations by assuming that the frequency of interest is much higher than $1/2\pi R_f C_f$ and that $R_o << R_f$. The factor $sC_{in}R_f/(1 + sR_fC_f)$ in front of the unity-feedback loop in Fig. 3-7(b) represents the ideal transfer function of the front-end amplifier, while the non-ideal dynamics are captured in the feedback loop part of the diagram. The ideal transfer function indicates that the low-frequency cutoff (highpass type) is given by $1/2\pi C_f R_f$. At the frequency greater than $1/2\pi R_f C_f$, the magnitude of the ideal closed-loop gain of the front-end amplifier is given by $A_m = C_{in}/C_f$. The loop transmission of the front-end amplifier is given by

$$\begin{aligned} L(s) &= \frac{C_f}{C_{in} + C_f + C_{p,in}} \cdot \frac{A(s)}{1 + sR_o\,(C_f + C_{p,out})} \\ &= \frac{C_f}{C_{in} + C_f + C_{p,in}} \cdot \frac{A_{OL}}{(1 + s/p_1)(1 + sR_o\,(C_f + C_{p,out}))} \end{aligned} \tag{3.4}$$

where we model the transfer function of A_1 as $A(s) = A_{OL}/(1 + s/p_1)$, where A_{OL} is the open-loop gain of A_1 and p_1 is the pole associated with the node $V_{out,OTA}$ of Fig. 3-5(b) due to the high impedance at this node. The loop transmission in (3.4) captures the loop dynamic, which determines the stability of the front-end amplifier. Fortunately, an explicit frequency compensation scheme is not needed since the desired closed-loop gain of 40 dB results in an attenuation of approximately 40 dB in the loop gain compared to that in a unity-gain amplifier case. This implicit reduce-gain compensation successfully stabilizes the amplifier. The attenuation in the loop gain is captured by the factor $C_f/(C_{in} + C_f + C_{p,in})$ in $L(s)$. In this design, we would like the closed-loop gain of the front-end amplifier, A_m, to be 40 dB (100), thus we have to make $C_{in} = 100C_f$. Because $C_{p,in}$ is just a parasitic capacitance at the input of the amplifier A_1, its value is less than the explicit capacitor C_f. Therefore, the factor

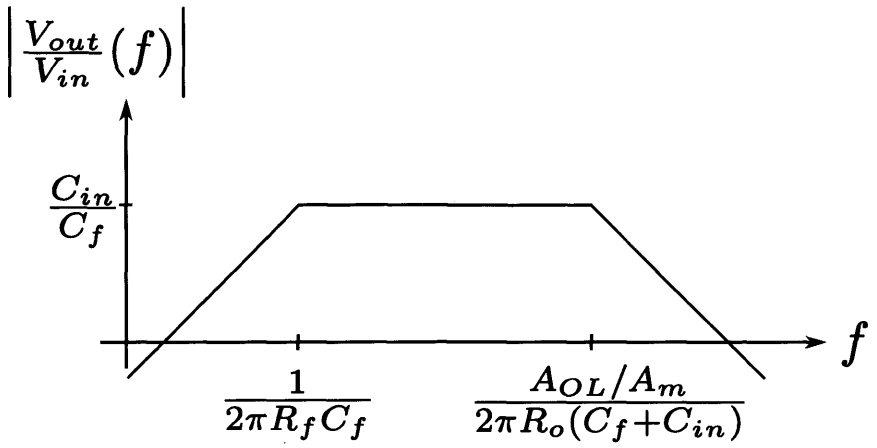


Figure 3-8: Bode magnitude plot of the front-end amplifier's transfer function.

$C_f/(C_{in}+C_f+C_{p,in})$ in $L(s)$ can be approximated as $1/A_m = 0.01$.

Next, let's determine the non-ideal transfer function of the front-end amplifier. From the block diagram of Fig. 3-7(b), we can write the transfer function of the front-end amplifier as

$$\frac{V_{out}}{V_{in}}(s) = -\frac{sC_{in}R_f}{1+sC_fR_f} \cdot \frac{A(s)/A_m}{1+A(s)/A_m} \cdot \frac{1}{1+sR_o(C_f+C_{in})/(A(s)/A_m)} \tag{3.5}$$

If we assume that $A(s) > A_m$ and that $A(s)$ can be approximated as A_{OL} at all the frequency of interest, (3.5) can be approximated as

$$\frac{V_{out}}{V_{in}}(s) = -\frac{sC_{in}R_f}{1+sC_fR_f} \cdot \frac{1}{1+sR_o(C_f+C_{in})/(A_{OL}/A_m)}. \tag{3.6}$$

It should be emphasized that (3.6) is just an approximation, since in reality the current in the OTA can be reduced according to the adaptive biasing scheme such that the magnitude of $A(s)$ drops significantly at higher frequency, and thus the assumptions that we have made are no longer valid. Nevertheless, (3.6) still provides useful insights into the operation of the front-end amplifier.

The Bode magnitude plot of the transfer function in (3.6) is shown in Fig. 3-8. From this plot, we can see that the midband gain of the front-end amplifier is C_{in}/C_f, while the low-frequency and high-frequency cutoff are $1/2\pi R_f C_f$ and $(A_{OL}/A_m)/2\pi R_o(C_f+C_{in})$ respectively. Since R_f is the resistance of the pseudoresis-

tor element and is very large, the low-frequency cutoff can be made lower than 1 Hz even with a small value of C_f. As mentioned earlier, the high-frequency cutoff of the front-end amplifier should be made as large as possible such that the bandwidth of the overall neural amplifier is determined by that of the bandpass filter stage. This is accomplished by the use of the class-AB output buffer to minimize R_o of the amplifier A_1 without consuming too much quiescent current.

Noise Analysis

The noise introduced by the front-end amplifier is of critical importance since it is the first stage in the signal processing chain. In this section, we will perform a detailed noise analysis of the front-end amplifier to understand how to minimize it while keeping the overall power consumption of the front-end amplifier low. From the feedback block diagram in Fig. 3-7(a), we can calculate the input-referred noise of the front-end amplifier by referring the input-referred noise of A_1, $\overline{v_{n,A1}^2}$, to the input V_{in}. Therefore, the input-referred noise of the front-end amplifier, $\overline{v_{n,amp}^2}$, is given by

$$\overline{v_{n,amp}^2} = \left| \frac{1 + sR_f(C_{in} + C_f + C_{p,in})}{sC_{in}R_f} \right|^2 \cdot \overline{v_{n,A1}^2} \tag{3.7}$$

At our frequency of interest ($f > 1/2\pi R_f C_f$), the expression in (3.7) can be simplified to

$$\overline{v_{n,amp}^2} = \left(\frac{C_{in} + C_f + C_{p,in}}{C_{in}} \right)^2 \cdot \overline{v_{n,A1}^2}. \tag{3.8}$$

To minimize the front-end amplifier's input-referred noise in (3.8), we need to minimize $\overline{v_{n,A1}^2}$, and also maximize C_{in} relative to C_f and $C_{p,in}$. Making the closed-loop gain of the front-end amplifier very large by making $C_{in} >> C_f$ would help minimize the input-referred noise of the front-end amplifier. However, making the front-end amplifier's gain too large is not practical because large LFPs may saturate the output of the front-end amplifier resulting in information lost. Therefore we limit the gain of the front-end amplifier to 40 dB. Additional gain is provided by the programmable-gain amplifier after the LFPs and spikes have been separated by the bandpass filter. It is important to note that $C_{p,in}$ must be kept small relative to C_{in}, otherwise, it can

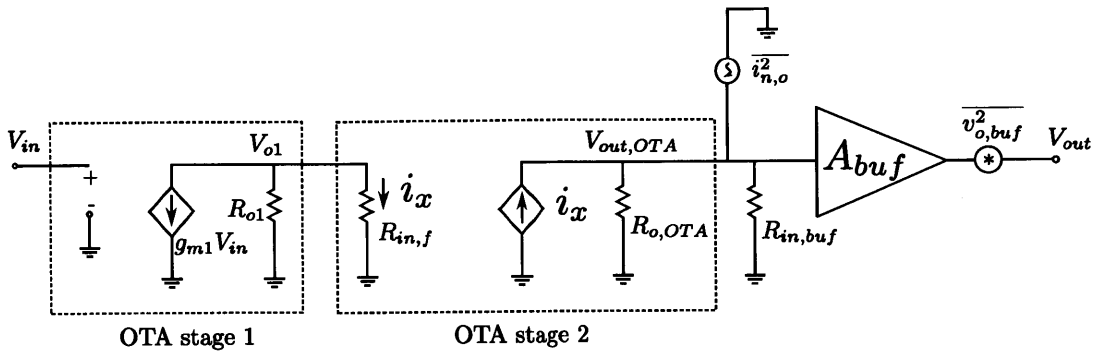


Figure 3-9: Small-signal diagram for the noise analysis of the amplifier A_1.

degrade $\overline{v_{n,amp}^2}$. As a result, making the input transistors M_1 and M_2 too large to minimize 1/f noise can adversely affect the overall input-referred noise of the neural amplifier.

To perform the noise analysis of the amplifier A_1, we use a small-signal diagram of Fig 3-9. In Fig 3-9, we model the OTA as a two-stage amplifier. The first stage of the OTA is modeled as a voltage amplifier with an infinite input resistance and the output resistance of R_{o1}. The input of the first stage OTA is V_{in} while the output is V_{o1}. The output voltage V_{o1} of the first stage corresponds to the voltage difference between the source nodes of M_9 and M_{10}. The first stage of the OTA consists of M_{b1}, M_{b2}, and M_1-M_8 of the circuit in Fig. 3-5(b). The second stage of the OTA is modeled as a transimpedance amplifier with an input resistance of $R_{in,f}$ and the output resistance of $R_{o,OTA}$. The input signal of the second stage OTA is the current i_x, while the output signal is the voltage $V_{out,OTA}$. The input current i_x corresponds to the differential current between M_9 and M_{10} of Fig. 3-5(b). The second stage of the OTA consists of the transistors M_9-M_{14} of the circuit in Fig. 3-5(b). The class-AB output buffer is modeled as a voltage amplifier with a gain of A_{buf}, and a finite input resistance of $R_{in,buf}$. The output resistance R_{o1} of the first stage OTA corresponds to the resistance looking into the drains of M_3 and M_5 (or M_4 and M_6). A simple analysis yields

$$R_{o1} \approx r_{o1}(g_{s3}r_{o3})||r_{o7}(g_{s5}r_{o5}) \tag{3.9}$$

where r_{oi} and g_{si} are the Early Effect resistance and the source admittance of the

transistor M_i respectively. The resistance $R_{in,f}$ is the resistance looking into the second stage of the OTA, which corresponds to the resistance looking into the source of M_9 (or M_{10}). Therefore,

$$R_{in,f} = 1/g_{s9}. \tag{3.10}$$

The circuit noise of A_1 is modeled as the current noise generator, $\overline{i_{n,o}^2}$, injected into the output node $V_{out,OTA}$ and the voltage noise source, $\overline{v_{o,buf}^2}$, in series with the output node V_{out}. These current noise and voltage noise sources, whose values will be calculated next, account for the noise contributions of all the transistors in A_1. The current noise generator $\overline{i_{n,o}^2}$ accounts for the noise contributions from M_1-M_{15}, M_{17}, and the two current sources I_{buf}. Note that each current source I_{buf} must be implemented with a transistor, and thus it contributes noise just like a regular transistor. The voltage noise generator $\overline{v_{o,buf}^2}$ accounts for the noise contributions from the output transistors M_{16} and M_{18}. Note that, due to the symmetry of the circuit, the noise contributions from M_{b1}-M_{b8} to the output node V_{out} are zero, and thus can be ignored.

For the noise analysis that follows, we model the noise of each transistor in the circuit of Fig. 3-5(b) as a current noise generator connected between its drain and source terminals. Let's denote the current noise generator of the transistor M_i as $\overline{i_{n,i}^2}$, $i \in \{1, ..., 18\}$, and the current noise generator of each current source I_{buf} as $\overline{i_{n,buf}^2}$. To calculate $\overline{i_{n,o}^2}$, we calculate and combine the current noise contributions from all the transistors to the node $V_{out,OTA}$, while it is being held at an AC ground. Let $\alpha = R_{o1}/(R_{o1} + R_{in,f})$ be the current division ratio at the source node of M_9 (or M_{10}). Intuitively, α represents the ratio of the incremental input current that can flow to the OTA's output node $V_{out,OTA}$. Since the current noise generators $\overline{i_{n,i}^2}$ are uncorrelated to each other, the output current noise $\overline{i_{n,o}^2}$ can be expressed as

$$\begin{aligned}
\overline{i_{n,o}^2} &= \alpha^2(\overline{i_{n,1}^2} + \overline{i_{n,2}^2} + \overline{i_{n,7}^2} + \overline{i_{n,8}^2}) + (1-\alpha)^2(\overline{i_{n,9}^2} + \overline{i_{n,10}^2}) \\
&\quad + \overline{i_{n,13}^2} + \overline{i_{n,14}^2} + 2\overline{i_{n,buf}^2} \\
&= 2\alpha^2(\overline{i_{n,1}^2} + \overline{i_{n,7}^2}) + 2(1-\alpha)^2\overline{i_{n,9}^2} + 2\overline{i_{n,13}^2} + 2\overline{i_{n,buf}^2}.
\end{aligned} \tag{3.11}$$

Every transistor in the amplifier A_1 is designed to operate in saturation, we can use a noise model for a MOS transistor that is valid in both subthreshold and above-threshold operations [21]

$$\overline{i_{n,i}^2} = \frac{8}{3}kTg_{m,i} + \frac{K_i g_{mi}^2}{W_i L_i C_{ox}} \frac{1}{f} \tag{3.12}$$

where k is the Boltzmann constant, K_i is the 1/f noise coefficient of the i^{th} transistor, and C_{ox} is the oxide capacitance per unit area. K_i takes the value of K_p if M_i is a PFET, while it takes the value of K_n if M_i is an NFET. The first term in (3.12) represents the thermal noise component while the second term represents the 1/f noise component of the transistor M_i. From Fig. 3-5(b), since I_{buf} is chosen just to be enough to bias the gate-to-source voltages of M_{15} and M_{17}, its value can be made much smaller than the current I_b in M_{13} and M_{14}. In our design, the value of I_{buf} is chosen to be 20 nA, while the lowest value of I_b is greater than 100 nA. Therefore, the g_m's of the transistors that make I_{buf} current sources are much smaller than the g_m's of M_1-M_{14}, and thus their noise contributions can be ignored. We can write the expression for $\overline{i_{n,o}^2}$ as

$$\begin{aligned}\overline{i_{n,o}^2} &= \frac{16kT}{3}\left[\alpha^2 g_{m1} + \alpha^2 g_{m7} + (1-\alpha)^2 g_{m9} + g_{m13}\right] \\ &+ \left[2\alpha^2\left(\frac{K_p g_{m1}^2}{W_1 L_1 C_{ox}} + \frac{K_n g_{m7}^2}{W_7 L_7 C_{ox}}\right) + 2(1-\alpha)^2 \frac{K_n g_{m9}^2}{W_9 L_9 C_{ox}}\right. \\ &\left. + 2\frac{K_p g_{m13}^2}{W_{13} L_{13} C_{ox}}\right] \frac{1}{f}\end{aligned} \tag{3.13}$$

The voltage noise generator at the output of the class-AB output buffer, $\overline{v_{o,buf}^2}$, can be expressed as

$$\overline{v_{o,buf}^2} = \frac{\overline{i_{n,16}^2} + \overline{i_{n,18}^2}}{(g_{s16} + g_{s18})^2}. \tag{3.14}$$

To find the input-referred noise of A_1 due to $\overline{i_{n,o}^2}$, we divide the expression in (3.13) by the transconductance of the OTA. Similarly, to find the input-referred noise of A_1 due to $\overline{v_{o,buf}^2}$, we divide the expression in (3.14) by the voltage gain from V_{in} to V_{out}.

The transconductance of the OTA is given by

$$G_m = g_{m1} \cdot \frac{R_{o1}}{R_{o1} + R_{in,f}} = \alpha g_{m1} \tag{3.15}$$

and the voltage gain from V_{in} to V_{out} is given by

$$\begin{aligned} A &= g_{m1} \cdot \frac{R_{o1}}{R_{o1} + R_{in,f}} \cdot (R_{o,OTA} || R_{in,buf}) \cdot A_{buf} \\ &\approx G_m \left(R_{o,OTA} || R_{in,buf}\right) \end{aligned} \tag{3.16}$$

where $A_{buf} = (g_{m16} + g_{m18})/(g_{s16} + g_{s18})$ is the gain of the class-AB output buffer whose value is less than but close to 1. The total input-referred noise of A_1 can be calculated from

$$\begin{aligned} \overline{v_{n,A1}^2} &= \frac{1}{G_m^2}\overline{i_{n,o}^2} + \frac{1}{A^2}\overline{v_{o,buf}^2} \\ &= \frac{1}{G_m^2}\overline{i_{n,o}^2} + \frac{\overline{i_{n,16}^2} + \overline{i_{n,18}^2}}{G_m^2 \left(g_{s16} + g_{s18}\right)^2 \left(R_{o,OTA} || R_{in,buf}\right)^2} \\ &\approx \frac{\overline{i_{n,o}^2}}{G_m^2}. \end{aligned} \tag{3.17}$$

The approximation in (3.17) is based on the fact that $(g_{s16} + g_{s18})(R_{o,OTA} || R_{in,buf}) \gg 1$. Note that $R_{o,OTA}$ is very high due to cascoding and $R_{in,buf}$ is also high since it is the parallel combination of the output resistances of the transistors that make the current sources I_{buf}. Thus $R_{o,OTA} || R_{in,buf}$ is already greater than $r_{o16} || r_{o18}$. Since we know that $(g_{s16}+g_{s18})(r_{o16} || r_{o18}) \gg 1$, then $(g_{s16}+g_{s18})(R_{o,OTA} || R_{in,buf}) \gg 1$. Therefore, the noise contribution from the class-AB output buffer is negligible. Combining (3.13)-

(3.17), we can write $\overline{v_{n,A1}^2}$ as

$$\begin{aligned}\overline{v_{n,A1}^2} &= \frac{8kT}{3g_{m1}}\left[1+\frac{g_{m7}}{g_{m1}}+\left(\frac{1-\alpha}{\alpha}\right)^2\cdot\frac{g_{m9}}{g_{m1}}+\frac{1}{\alpha^2}\frac{g_{m13}}{g_{m1}}\right] \\ &+\left[\frac{2K_p}{W_1L_1C_{ox}}+\frac{2K_n}{W_7L_7C_{ox}}\left(\frac{g_{m7}}{g_{m1}}\right)^2+\left(\frac{1-\alpha}{\alpha}\right)^2\frac{2K_n}{W_9L_9C_{ox}}\left(\frac{g_{m9}}{g_{m1}}\right)^2\right. \\ &\left.+\frac{1}{\alpha^2}\frac{2K_p}{W_{13}L_{13}C_{ox}}\left(\frac{g_{m13}}{g_{m1}}\right)^2\right]\frac{1}{f}. \end{aligned} \tag{3.18}$$

The first and second terms on the right hand side of (3.18) represent the input-referred thermal noise component and the input-referred $1/f$ noise component of the amplifier A_1 respectively. To minimize the noise contributions from M_9 and M_{10}, we need to maximize the ratio α. We maximize α by adding the cascode transistors M_3, M_4, M_5, and M_6 to the drains of M_1, M_2, M_7 and M_8 respectively. Cascoding makes the admittance looking into the source terminals of M_9 and M_{10} much larger than the admittance looking into the drains of M_3, M_5 and M_4, M_6 even when the current in M_9 and M_{10} is a small fraction of the current in M_3, M_5 and M_4, M_6. Quantitatively, we make R_{o1} in (3.9) much larger than $R_{in,f}$ in (3.10). As a result, α achieves a value close to 1 and (3.18) is reduced to

$$\begin{aligned}\overline{v_{n,A1}^2} &= \frac{8kT}{3g_{m1}}\left[1+\frac{g_{m7}}{g_{m1}}+\frac{g_{m13}}{g_{m1}}\right]+\left[\frac{2K_p}{W_1L_1C_{ox}}+\frac{2K_n}{W_7L_7C_{ox}}\left(\frac{g_{m7}}{g_{m1}}\right)^2\right. \\ &\left.+\frac{2K_p}{W_{13}L_{13}C_{ox}}\left(\frac{g_{m13}}{g_{m1}}\right)^2\right]\frac{1}{f}. \end{aligned} \tag{3.19}$$

To minimize both the thermal noise and $1/f$ noise components of A_1 for a given bias current, we need to maximize g_{m1}, while minimizing g_{m7} and g_{m13} relative to g_{m1}. The transconductance g_{m1} is maximized for a given bias current if M_1 and M_2 operate deep in subthreshold. To operate M_1 and M_2 deep in subthreshold, we size them with a large W/L ratio (200 μm/1.2 μm in this design). To minimize g_{m13} relative to g_{m1}, first we scale the current in M_{13} and M_{14} such that it is only a small fraction of the current in M_1 and M_2. In our design, the current in M_{13} and M_{14} is set to $1/5^{th}$

of the current in M_1 and M_2. This current scaling scheme ensures that most of the supply current in the OTA is consumed in the input differential pair transistors M_1 and M_2, and not in the folded branch (M_9 - M_{14}), in which the higher bias current would increase both the input-referred noise and the total power consumption of A_1. The distribution of the current in the OTA is achieved with the help of the bias circuit formed by the transistors M_{b1}, M_{b2}, M_{b3}, M_{b4}, M_{b7}, M_{b8} and M_5-M_8. As a result, g_{m13} is made small relative to g_{m1}, minimizing both the noise contributions from M_{13} and M_{14} and the total power consumption of A_1. To further reduce g_{m13} relative to g_{m1}, we size M_{13} and M_{14} with a small W/L ratio such that they operate in strong inversion where transistors exhibit smaller g_m for a given bias current (a small g_m/I_D ratio). The transistors M_7 and M_8 carry a higher bias current than that in M_1 and M_2, and thus can contribute a significant amount of noise. To minimize the noise contributions from M_7 and M_8, we need to minimize g_{m7} by sizing M_7 and M_8 with a small W/L ratio (12 μm/64 μm) such that they operate deep in the strong inversion region with a large gate overdrive voltage.

The 1/f noise power per unit bandwidth of a FET is proportional to its g_m^2 and inversely proportional to its gate area ($W \times L$). Because the input transistors' g_m is the highest among the transistors in the OTA, the only way to minimize their 1/f noise contributions without using excessively large input transistors is to use pFETs since they have the smallest 1/f noise coefficient ($K_p \approx 0.1K_n$ in our process). Furthermore, the transistors that can contribute significant 1/f noise (M_1, M_2, M_7 and M_8) due to their relatively high g_m's are made with large gate areas. The sizes of the transistors in the OTA along with their intended regions of operation are summarized in Table 3.1.

Due to the high gain of the front-end amplifier, its input-referred noise dominates the input-referred noise of the overall neural amplifier. Therefore, controlling the front-end amplifier's input-referred noise provides the most effective way of controlling the input-referred noise of the overall neural amplifier. From (3.19), due to the fixed ratios of currents in M_7 and M_1 and in M_{13} and M_1, the ratios g_{m7}/g_{m1} and g_{m13}/g_{m1} are relatively constant across the OTA's bias current. In subthreshold, g_{m1} is proportional to M_1's bias current, and thus the thermal noise component in (3.19)

Table 3.1: Transistor sizings of the OTA in the amplifier A_1

Devices	W/L (μm)	Operating region
M_1, M_2	20 × 10/1.2	subthreshold
M_3, M_4	10 × 6/0.36	subthreshold
M_5, M_6	5 × 10/0.64	subthreshold
M_7, M_8	6 × 2/64	strong inversion
M_9, M_{10}	4 × 4/0.64	subthreshold
M_{11}, M_{12}	4 × 6/0.2	subthreshold
M_{13}, M_{14}	2 × 2/10	strong inversion
M_{b1}, M_{b2}	10 × 5/2.4	moderate inversion
M_{b3}, M_{b4}	5/2.4	moderate inversion
M_{b6}	10/0.64	subthreshold
M_{b7}	2/64	strong inversion

is inversely proportional to M_1's bias current. However, the 1/f noise component in (3.19) is invariant to M_1's bias current. By keeping the bandwidth of the neural amplifier constant and ensuring that the thermal noise component dominates the overall noise of the OTA, we can control the overall input-referred noise of the neural amplifier by controlling the bias current in M_1 and M_2; we increase the bias current in M_1 and M_2 to reduce the amplifier's input-referred noise, and vice versa. The bias current of the OTA is controlled by a 4-bit binary current DAC which is represented as the variable current source I_b in Fig. 3-5b).

3.2.2 Bandpass Filter

Figure 3-10(a) shows the schematic of the bandpass filter used in this design. The bandpass filter can be configured with two recording settings: i) a spike recording setting (350 Hz - 12 kHz) and ii) an LFP recording setting (< 1 Hz - 300 Hz). The choice of the recording setting is controlled by the signal ϕ_1; when $\phi_1 = 1$, the filter is in the spike recording setting, and when $\phi_1 = 0$, the filter is in the LFP recording setting. Let's denote the upper and lower -3 dB cutoff frequencies of the bandpass filter as f_h and f_l respectively. To set the value of f_h, we use the combination of the unity-gain connected G_{m2}-OTA and the load capacitance C_2. The cutoff frequency f_h can be expressed as $f_h = G_{m2}/2\pi C_2$ where G_{m2} is the effective transconductance of

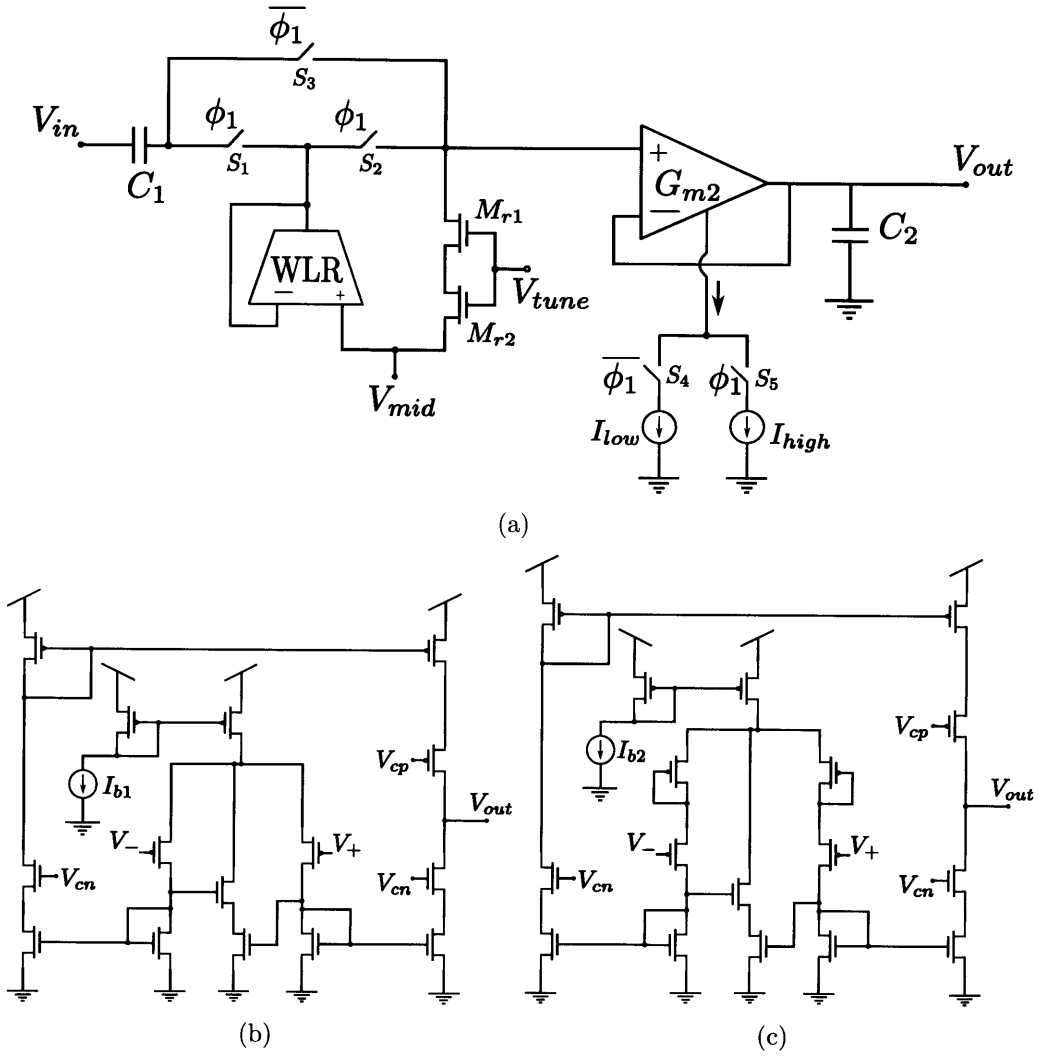


Figure 3-10: (a) Schematic of the bandpass filter. (b) Schematic of the G_{m2}-OTA. (c) Schematic of the WLR-OTA.

the G_{m2}-OTA, which is a function of the OTA's bias current. In the spike recording setting ($\phi_1 = 1$), the switch S_5 is closed and the G_{m2}-OTA is biased with the current I_{high} such that $f_h(\phi_1 = 1) = G_{m2}(I_{high})/2\pi C_2 = 12$ kHz. Similarly, in the LFP recording setting ($\phi_1 = 0$), the switch S_4 is closed and G_{m2}-OTA is biased with I_{low} such that $f_h(\phi_1 = 0) = G_{m2}(I_{low})/2\pi C_2 = 350$ Hz.

Setting the cutoff frequency f_l is slightly more subtle. To set f_l in the spike recording setting, we use a combination of C_1 and the unity-gain connected WLR-OTA [53]. However, due to the difficulties of biasing a WLR-OTA at very low bias current to achieve $f_l < 1$ Hz, we simply use the series combination of M_{r1} and M_{r2} as a large resistor to set f_l and to provide a DC path to the positive input terminal of G_{m2}-OTA in the LFP recording setting. Let R_p denote the effective resistance of the series combination of M_{r1} and M_{r2} and let $G_{m,WLR}$ denote the effective transconductance of the WLR-OTA. For now, let's assume that the voltage V_{tune} is set such that $R_p >> 1/G_{m,WLR}$. In the spike recording setting ($\phi_1 = 1$), the switches S_1 and S_2 are closed and the WLR-OTA appears in parallel with R_p. Since $R_p >> 1/G_{m,WLR}$, the effective resistance $1/G_{m,WLR}$ dominates the parallel combination. As a result, the cutoff frequency f_l in the spike recording setting is approximated as $f_l = G_{m,WLR}/2\pi C_1$. The bias current of the WLR-OTA is set such that $f_l = 350$ Hz. In the LFP recording setting ($\phi_1 = 0$), the switches S_1 and S_2 are open while the switch S_3 is closed. The WLR-OTA is disconnected from the signal path and the capacitor C_1 appears in series with R_p through the switch S_3. In this case, the combination of C_1 and R_p determines the cutoff frequency f_l. By setting $V_{tune} = V_{mid}$, the effective resistance R_p is very large such that $f_l = 1/2\pi R_p C_1 < 1$Hz. Note that the drain and source terminals of M_{r1} and M_{r2} are approximately at V_{mid} while the bulk terminals of the two transistors are at ground. The resulting Body Effect helps minimize the subthreshold leakage current through M_{r1} and M_{r2} and allows a low value of f_l.

The schematics of G_{m2}-OTA and WLR-OTA are shown in Fig. 3-10(b) and 3-10(c) respectively. To achieve low cutoff frequencies of 300 Hz and 12 kHz with reasonable bias currents in the OTAs and small capacitances C_1 and C_2, we utilize G_m reduction

techniques described in [54] or in [53]. Bump-linearization and source degeneration techniques are used in the WLR-OTA to achieve a very low cutoff frequency of 300 Hz. To achieve a 12-kHz cutoff frequency in the G_{m2}-OTA, only the bump-linearization technique is deployed.

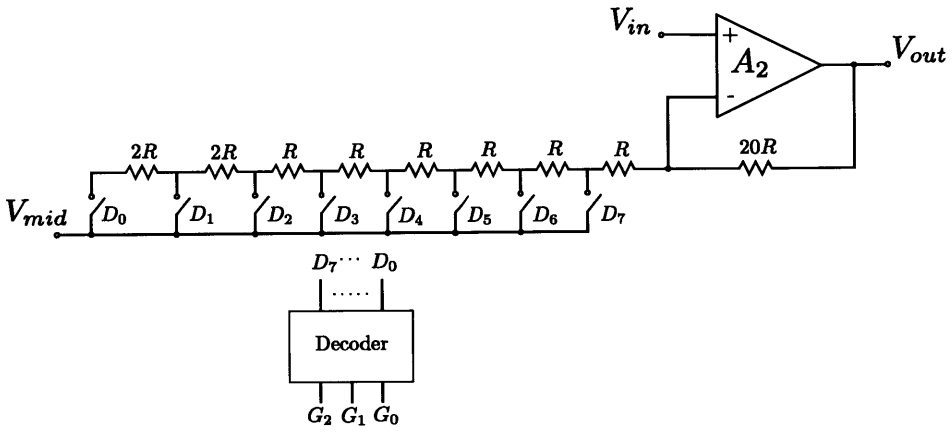
3.2.3 Programmable Gain Amplifier

Figure 3-11(a) shows the schematic of the programmable-gain amplifier. To achieve a large output signal swing with low distortion, we use a non-inverting amplifier topology with linear resistors in the feedback path. The gain of the programmable-gain amplifier can be programmed to any of the eight values and is given by $A_{v2} = 1+\frac{20R}{R_v(D_i)}$, $i \in \{0, ..., 7\}$, where $R_v(D_i)$ is the total resistance seen between the negative input terminal of A_2 and the node V_{mid} when the switch D_i is closed. The digital decoder ensures that only one of the switches D_i's can be closed at a given time depending on the decoder's inputs G_2, G_1, and G_0. The values of the gain A_{v2} for every combination of G_2, G_1, and G_0 are tabulated in Table 3.2.

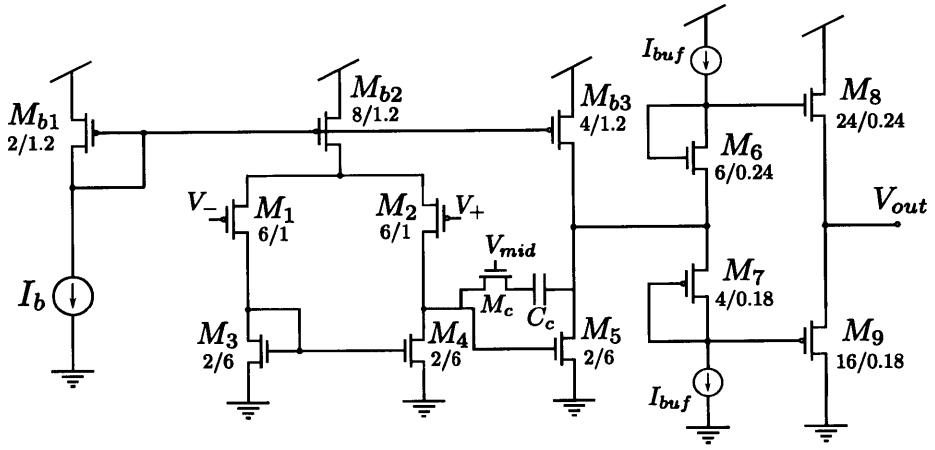
The schematic of the amplifier A_2 is shown in Fig. 3-11(b). We use a standard two-stage design with Miller compensation [51,53]. The capacitor C_c acts as a Miller capacitor and the transistor M_c, biased in the linear region, acts as a resistor to eliminate the right-half-plane zero. We design A_2 to be unity-gain stable such that it has ample stability margins even when configured in the low closed-loop gain settings. The class-AB output buffer is included to drive the resistive load (the feedback resistors) at the output of A_2.

Table 3.2: Gains of the programmable-gain amplifier

G_2	G_1	G_0	switch closed	Gain (dB)
0	0	0	D_0	9.5
0	0	1	D_1	10.8
0	1	0	D_2	12.7
0	1	1	D_3	14
1	0	0	D_4	15.6
1	0	1	D_5	17.7
1	1	0	D_6	20.8
1	1	1	D_7	26.4



(a)



(b)

Figure 3-11: (a) Schematic of the programmable-gain amplifier. (b) Schematic of the amplifier A_2.

3.3 Neural Signal Digitization

After amplification, the neural signals are converted into digital format by analog-to-digital converters (ADCs) to facilitate the communication between the 32-channel neural recording IC and the on-board FPGA. Due to power constraint of an implantable recording system, the power overhead per channel due to signal digitization should also be kept as small as possible. In this 32-channel neural recording IC, one ADC is shared by four neural amplifiers in a recording module to save silicon area. In this section, we present the design of an energy-efficient analog-to-digital converter that is used in the 32-channel neural recording IC, along with the design of an analog multiplexer used for multiplexing the outputs of neural amplifiers into the ADC. To minimize power consumption, duty cycling technique is utilized to turn on various circuits only when they are needed.

3.3.1 ADC Basic

Before we describe the design of our ADC, let's review a few basic concepts that are important to understand the motivation behind our ADC's specifications.

ADC's Quantization Noise

An N-bit ADC is a a circuit building block that converts a continuous time signal V_{in} into a series of N-bit digital output words $\{b_{N-1},...,b_0\}$. For an ideal ADC, the relationship between the input signal V_{in} and the digital output word can be written as

$$V_{in} = V_{ref} \sum_{i=0}^{N-1} \frac{b_i}{2^{i+1}} + V_Q \tag{3.20}$$

where V_{ref} is a full-scale reference voltage, and V_Q is a quantization error due to the finite resolution of the ADC.

The relationship between the quantization error V_Q and the input signal V_{in} can be understood with an aid of a diagram in Fig. 3-12. In this diagram, an ideal digital-to-analog converter (DAC) converts the digital output of the ADC back to an analog

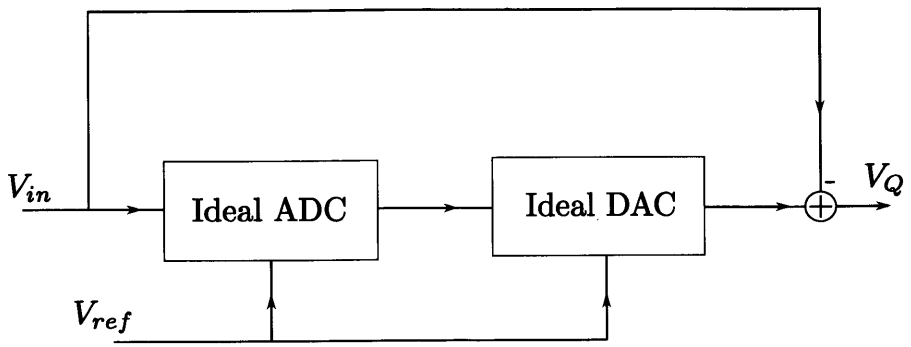


Figure 3-12: Circuit diagram illustrating the concept of quantization noise.

value. The difference between the output of the DAC and the input signal V_{in} is the quantization error V_Q. Due to this quantization error, we can think of an ADC as a unity-gain circuit that converts an analog value into a digital representation, while adding some noise to the signal. The noise due to the quantization error is called the quantization noise.

The quantization noise of the ADC limits the signal-to-noise ratio (SNR) of the ADC's output signal. Even an ideal ADC adds quantization noise to the ADC's output signal. For a real ADC, the error between the output of the ADC and the input is even worse since circuit's noises and imperfections also contribute to the error, resulting in a lower SNR of the converted signal. For an excellent treatment of ADC's SNR, please see [8]. To reduce the quantization noise, an ADC with a higher resolution (higher N) can be used, however, at the expense of higher complexity, larger silicon area, and higher power consumption. Due to the power and area requirements of the neural recording application, the ADC's resolution must be chosen carefully such that the ADC does not consume too much power or silicon area, while still being able to perform the conversion without adding significant amount of noise to the digitized signal.

To determine the needed precision of the ADC for our neural recording application, let's consider how much quantization noise that an N-bit ADC adds to the signal path. For an ideal N-bit ADC with a full-scale voltage of V_{ref}, the step size between the adjacent codes is $\Delta = V_{ref}/2^N$. The root-mean-square (rms) quantization noise can be approximated as $V_{Q,rms} = \Delta/\sqrt{12}$. This result is easy to understand using a

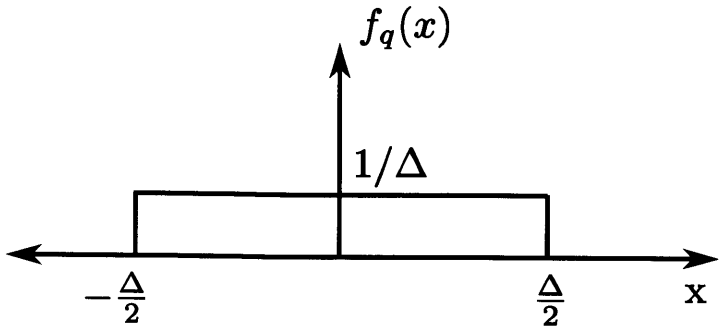


Figure 3-13: Probability density function of an ideal ADC's quantization noise.

probabilistic approach by assuming that the input signal V_{in} is varying very rapidly such that the quantization error V_Q due to the conversion is a uniform random variable with a value between $-\Delta/2$ and $\Delta/2$. The probability density function $f_q(x)$ of V_Q can be modeled as uniform in the range of $[-\Delta/2, \Delta/2]$ with an amplitude of $1/\Delta$ as shown in Fig. 3-13. The mean value of the quantization noise V_Q can be calculated as

$$\overline{V_Q} = \int_{-\infty}^{\infty} f_q(x) \cdot x \, \mathrm{d}x = \frac{1}{\Delta} \cdot \int_{-\Delta/2}^{\Delta/2} x \, \mathrm{d}x = 0. \tag{3.21}$$

Since V_Q is a zero-mean random variable, its variance can be calculated as

$$\overline{V_Q^2} = \int_{-\infty}^{\infty} f_q(x) \cdot x^2 \, \mathrm{d}x = \frac{1}{\Delta} \cdot \int_{-\Delta/2}^{\Delta/2} x^2 \, \mathrm{d}x = \frac{\Delta^2}{12}. \tag{3.22}$$

Thus, the rms quantization noise of the ideal ADC is $V_{Q,rms} = \sqrt{\overline{V_Q^2} - \overline{V_Q}^2} = \sqrt{\Delta^2/12} = \Delta/\sqrt{12}$.

ENOB and SNDR

For an ideal ADC, the quantization noise is the only noise source that affects the SNR of the digitized signal. The SNR of an ideal ADC is often characterized by assuming that the input signal is a full-scale sinusoidal signal (peak-to-peak amplitude of V_{ref}). For an ideal ADC with a full-scale sine wave input (with an amplitude of $V_{ref}/2$), the output signal power is $V_{ref}^2/8$, while the quantization noise power is

$\Delta^2/12 = V_{ref}^2/\left(12 \cdot 2^{2N}\right)$. The SNR of an ideal ADC can be expressed in dB as

$$\begin{aligned} \text{SNR} &= 10\log_{10}\left(\frac{\text{signal power}}{\text{quantization noise power}}\right) \\ &= 10\log_{10}\left(\frac{V_{ref}^2/8}{V_{ref}^2/\left(12 \cdot 2^{2N}\right)}\right) \\ &= 6.02N + 1.76 \text{ dB}. \end{aligned} \tag{3.23}$$

For a real ADC, the number of output bits (N) does not summarize the ADC's dynamic performance. Other circuit noises and nonidealities can worsen the SNR of the digitized signal compared to that in the ideal case when only the quantization noise is considered. To characterize the dynamic performance of a real ADC, we often use a metric called the signal-to-noise-and-distortion-ratio (SNDR). The SNDR is the ratio of the power between that of the fundamental component of the ADC's output signal and all other spectral components present in the output signal combined. The non-fundamental components stem from both the intrinsic circuit noise and the distortion caused by ADC's nonidealities. To obtain the SNDR, a rail-to-rail low-distortion sinusoidal signal is fed into the ADC's input. The digital output codes of the ADC are then collected and the fast-fourier transform (FFT) [43] is performed on the output samples to obtain a spectral plot. By performing an M-point FFT on M output samples and assuming that the fundamental frequency appears in bin m, the SNDR can be calculated from

$$\text{SNDR} = 10\log_{10}\left[\frac{A_m^2}{\sum_{i=1}^{m-1} A_i^2 + \sum_{i=m+1}^{M/2} A_i^2}\right] \tag{3.24}$$

where A_i is the amplitude of the spectral content in frequency bin i. The effective number of bits (ENOB) is just the SNDR expressed in bits rather than in dB. With an aid of equation (3.23) that relates the number of bits of an ideal ADC to the SNR, the ENOB can be expressed as

$$\text{ENOB} = \frac{\text{SNDR(in dB)} - 1.76}{6.02}. \tag{3.25}$$

The concepts of ENOB and SNDR will be used to characterize the dynamic performance of the ADC in Section 3.7.

3.3.2 ADC Design Considerations

Neural spikes exhibit frequency contents in the range of 300 Hz - 10 kHz, while LFPs exhibit frequency contents in the range of <1 Hz - 300 Hz. To sample these neural signals, the sampling speed requirement of the ADC is quite modest. Given the first-order roll-off of the neural amplifier, it is advantageous to sample the signals at slightly higher than the Nyquist frequency [43] to minimize aliasing due to the out-of-band noise. Since the output of each neural amplifier is bandlimited to 12 kHz, we choose the sampling rate per channel in the spike recording setting to be 31.25 kHz for convenience in generating the sampling signal from the external 10 MHz reference clock. To ease the design of the digital control logic block that controls the operation of the ADCs, the same sampling rate per channel is also used in the LFP recording setting. Therefore, the total sampling rate of each ADC is 125 kS/s because each ADC must digitize the neural signals from four neural amplifiers in a recording module.

Let's consider what ADC's resolution is needed for our application. First, let's consider a neural amplifier with a gain of 60 dB and an input-referred noise of 5 μV_{rms}. Note that this noise is just the intrinsic noise of the amplifier itself, and does not include the noise from neural background activity and the thermal noise of the high-impedance electrode. For an 8-bit ADC with a 1-V full-scale voltage, the rms quantization noise of such ADC is $\Delta/\sqrt{12} = 1/\left(2^8 \cdot \sqrt{12}\right) = 1.13\ mV_{rms}$. Thus, the value of the ADC's quantization noise when referred to the input of this 60 dB-gain amplifier is 1.13 μV_{rms}. Since the neural amplifier's intrinsic noise and the ADC's quantization noise are uncorrelated, the total noise when referred to the input of the amplifier is

$$V_{in,amp,rms} = \sqrt{(5\ \mu V_{rms})^2 + (1.13\ \mu V_{rms})^2} = 5.13\ \mu V_{rms}. \tag{3.26}$$

If the gain of the amplifier is reduced by half (gain of 500), the ADC's quantization noise referred to the input of the amplifier is 2.26 μV_{rms}, and thus the total noise referred to the input of the amplifier is 5.48 μV_{rms}. It can be seen that, even with 8-bit precision, the ADC's quantization noise hardly affects the overall noise of the signal chain provided that the neural amplifier's gain is high. In practice, the noise seen at the neural amplifier's input can be significantly higher than 5 μV_{rms} due to the neural background activity and the thermal noise from the high-impedance electrode. Therefore, an 8-bit ADC with a 1-V full-scale voltage (V_{ref}) should have sufficient resolution for the neural recording applications. Using the ADC with higher precision than 8 bits would not significantly improve the quality of the neural signals but only adds design complexity, power consumption, silicon area, and the amount of data that needs to be sent off-chip. Therefore, for our 32-channel neural recording IC, we choose to implement the ADCs using the successive-approximation-register (SAR) architecture due to its energy efficiency and small area when implemented at 8-bit resolution.

3.3.3 Basic Operation of the ADC

In this section, we describe the basic operation of our SAR ADC. The schematic of the ADC is shown in Fig. 3-14(a). The high-level topology of the ADC is similar to the one presented in [57]. The ADC consists of a comparator, a custom dynamic SAR logic, a switch network, a capacitor DAC array, and a bootstrapped reference switch. However, the comparator and the SAR logic are redesigned to achieve significant improvement in energy efficiency compared to the ADC presented in [57]. Since the ADCs in all the eight recording modules operate in parallel, we generate the ADC's clock and control signals in the centralized Digital Control Unit. These control signals are derived from an external 10 MHz clock and are common among all the ADCs on chip. The clock signal CK is the main clock of the ADC and is used for controlling the timing operation of the SAR logic while CK_s is used for registering the outputs of the comparator. The control signals SMP, $STRT$, EN are used by the SAR logic for sampling the input voltage V_{in}, initiating the conversion process, and duty

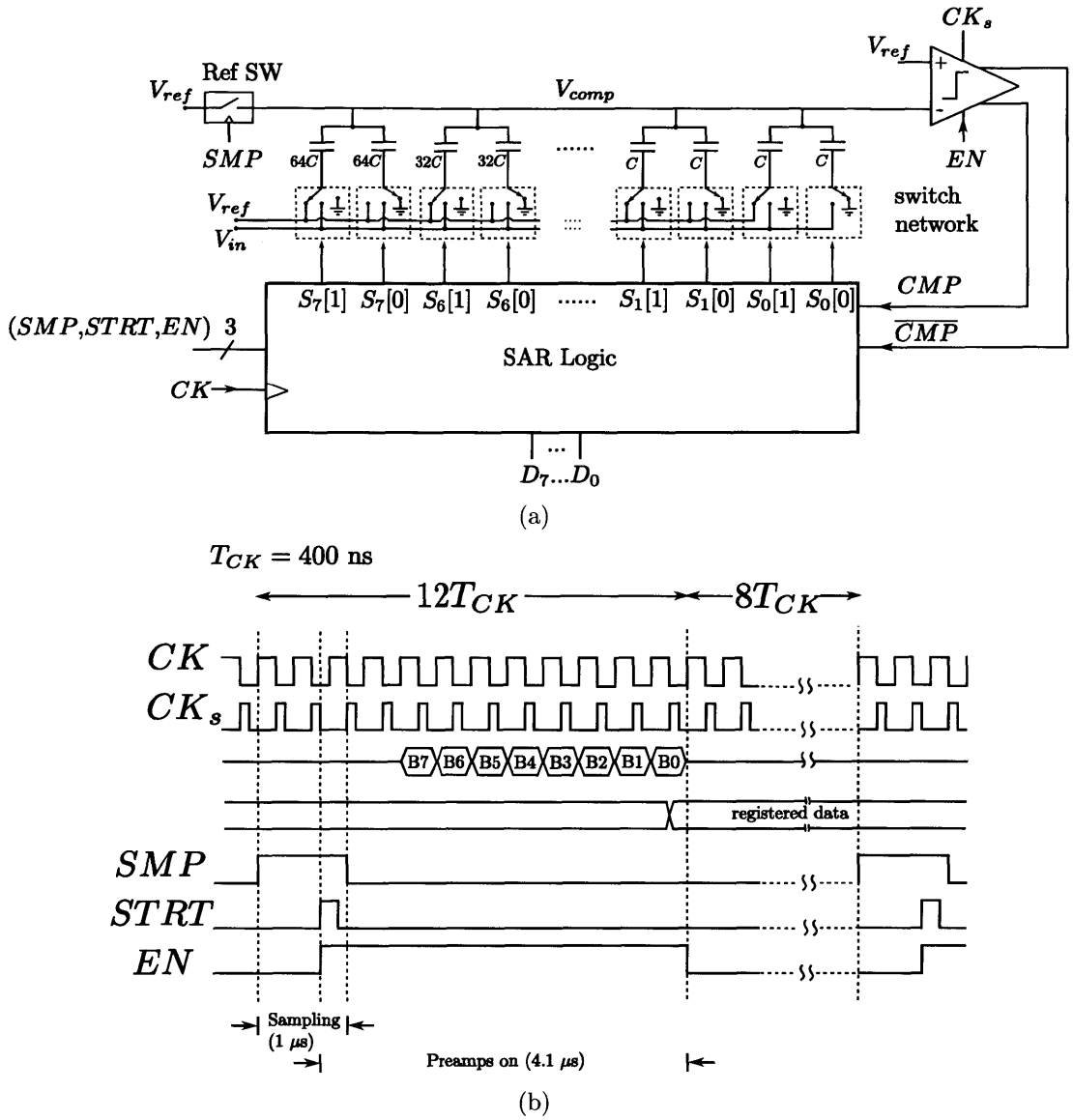


Figure 3-14: (a) Schematic of the SAR ADC used in this neural recording system. (b) Timing diagram of the ADC.

cycling the comparator to reduce power consumption respectively. The full-scale reference voltage of the ADC is $V_{ref} = 1$ V which is generated from an on-chip bandgap reference circuit. The split-capacitor approach presented in [19] is utilized to reduce the power consumed by the capacitor DAC array. Each of the binary-weighted capacitor in the array is divided in half while the two halves form a bit cell. The MSB bit cell or bit cell 7 (controlled by $S_7[1]$ and $S_7[0]$) consists of the left and right $64C$ capacitors, while the 2nd MSB bit cell or bit cell 6 consists of the left and right $32C$ capacitors, and so on. The unit capacitance of the capacitor DAC array in Fig. 3-14(a) is $C = 8$ fF.

The operation of the ADC can be described with an aid of the timing diagram shown in Fig. 3-14(b). During the sampling period of 1 μs (when SMP is high), the ADC's input voltage V_{in} is sampled onto the bottom plates of all the capacitors in the array while the reference voltage V_{ref} is sampled onto the common node V_{comp} through the bootstrapped reference switch. The $STRT$ signal is asserted a short moment before the sampling process ends (200 ns before the negative edge of SMP), initiating the SAR logic to start the successive approximation process. At the moment when the sampling period ends (SMP goes low), V_{ref} and V_{in} are disconnected from the top and bottom of the capacitor DAC array respectively. At this point, the voltage across all the capacitors in the capacitor DAC array is $V_{ref} - V_{in}$ while the voltage at the node V_{comp} is equal to V_{ref}. The SAR logic then immediately makes the first approximation by connecting the bottom plates of the left capacitors in all the bit cells to V_{ref}, while connecting the bottom plates of the right capacitors in all the bit cells to ground. The configuration of the capacitor DAC array at this point is as illustrated in Fig. 3-14(a). Effectively, half of the total capacitance in the DAC array is connected to V_{ref} while another half is connected to ground. Disconnecting the bottom plates of half of the capacitors in the DAC array from V_{in} and connecting them to V_{ref} add the voltage of value $\frac{1}{2}(V_{ref} - V_{in})$ to the node V_{comp}. Similarly, disconnecting the bottom plates of the other half of the capacitors in the DAC array from V_{in} and connecting them to ground subtracts the voltage of value $\frac{1}{2}V_{in}$ from the node V_{comp}. By the principle of superposition, the voltage at the node V_{comp} after the

first approximation by the SAR logic is $V_{comp} = V_{ref} + \frac{1}{2}(V_{ref} - V_{in}) - \frac{1}{2}V_{in} = \frac{3}{2}V_{ref} - V_{in}$.

The ADC then proceeds according to the successive approximation algorithm to determine all the eight digital bits that represent the input signal V_{in} from the most-significant-bit (MSB), D_7, down to the least-significant-bit (LSB), D_0. The successive approximation process is summarized in the flow graph of Fig. 3-15. A similar explanation for the SAR ADC using the flow graph technique can be found in [28]. In the first bit cycling period ($i = 1$ or period B7 in Fig. 3-14(b)), the ADC determines the value of the MSB (D_7). First, the voltage at the node V_{comp} is compared to V_{ref} by the comparator. If $V_{comp} < V_{ref}$, that means $V_{in} > V_{ref}/2$ and the output of the comparator CMP becomes 1, signifying that the MSB (D_7) is a 1. On the contrary, if $V_{comp} > V_{ref}$, that means $V_{in} < V_{ref}/2$ and CMP becomes 0, signifying that D_7 is a 0. The SAR logic then makes the next approximation based on the value of CMP by either adding or subtracting the voltage of value $V_{ref}/4$ to the voltage at the node V_{comp} by appropriately changing the configuration of the $64C$ capacitors in bit cell 7. If CMP is a 1, the SAR logic adds $V_{ref}/4$ to V_{comp} by switching the right $64C$ capacitor from ground to V_{ref}, while leaving the left $64C$ capacitor to remain at V_{ref}. However, if CMP is a 0, the SAR logic subtracts $V_{ref}/4$ from V_{comp} by switching the left $64C$ capacitor from V_{ref} to ground, while leaving the right $64C$ capacitor to remain at ground. This completes the first bit cycling period ($i = 1$ or the period B7 as labeled in Fig. 3-14(b)). The SAR logic then repeats the process for the next bit cycling periods ($i = 2, ..., 8$). This time, once the corresponding digital bit D_{8-i} has been determined, the SAR logic makes the next approximation by changing the configuration of the capacitors in bit cell $8 - i$ to either add or subtract the voltage of value $V_{ref}/2^{i+1}$ to or from V_{comp}. To add the voltage of value $V_{ref}/2^{i+1}$ to V_{comp}, the left capacitor in bit cell $8 - i$ is to remain at V_{ref}, while the right capacitor is switched from ground to V_{ref}. On the contrary, to subtract the voltage of value $V_{ref}/2^{i+1}$ from V_{comp}, the left capacitor in bit cell $8 - i$ is switched from V_{ref} to ground, while the right capacitor is to remain at ground. This process is repeated until the capacitors in bit cell 0 are properly reconfigured ($i = 8$) and all the digital bits have been determined.

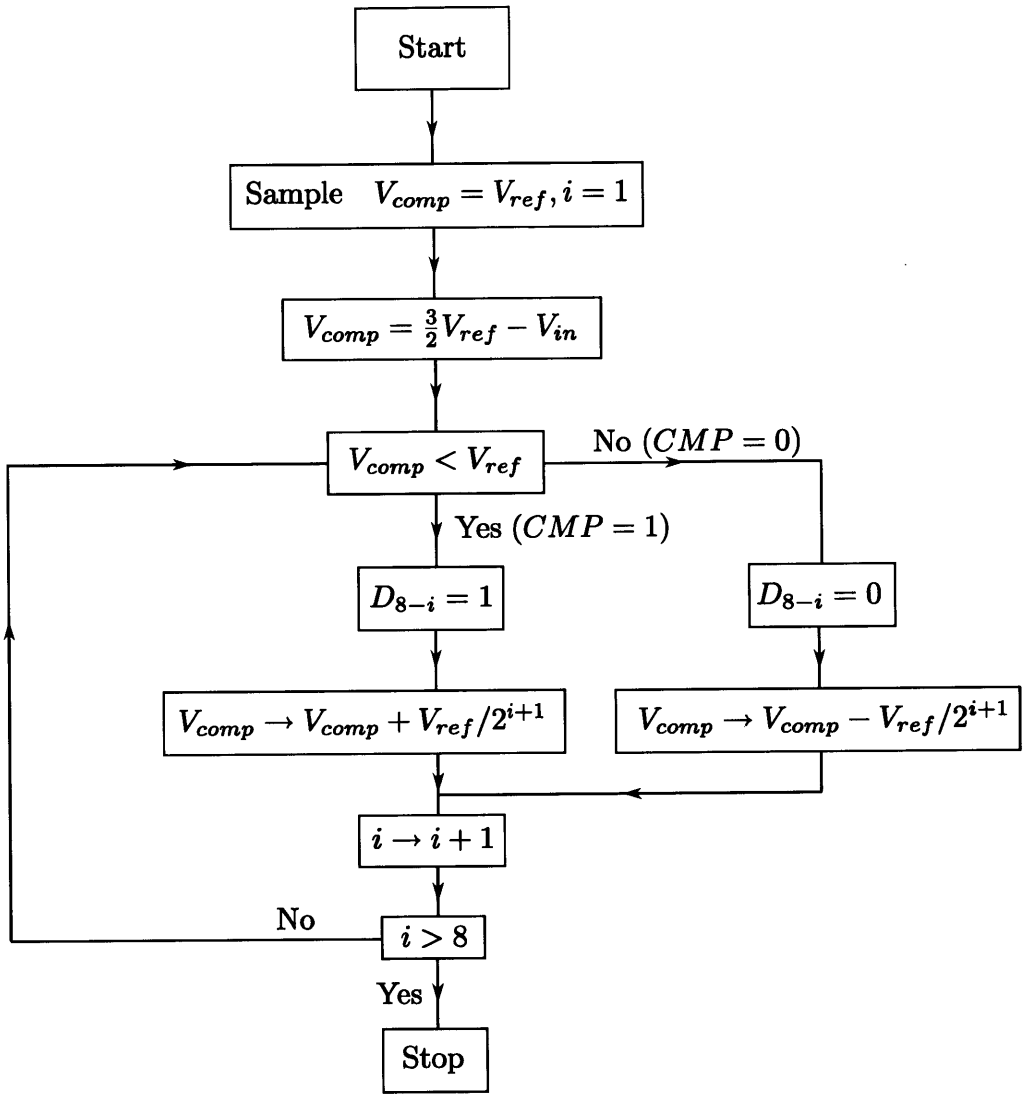


Figure 3-15: Flow graph illustrating the operation of the successive approximation ADC.

3.3.4 Circuit Implementations of the ADC

In this section, we describe the detailed implementations of the circuit building blocks of the ADC including the bootstrapped reference switch, the comparator, the custom SAR logic.

Bootstrapped Reference Switch

Once V_{ref} has been sampled onto the node V_{comp} and the reference switch has been opened at the end of the sampling phase, we must ensure that the reference switch remains open throughout the bit cycling periods. The ADC operates from a supply voltage of 1 V, which is the same as the reference voltage V_{ref}. However, depending on the input signal V_{in} at the sampling instant, the voltage at V_{comp} during the bit cycling periods can be either significantly higher or significantly lower than V_{ref} and the ADC's power supply voltage. For instance, if $V_{in} = 0$ V, during the first approximation by the SAR logic, the voltage at V_{comp} can reach $\frac{3}{2}V_{ref} - V_{in}$ which is 1.5 V. If $V_{in} = V_{ref}$, the voltage at V_{comp} after the first approximation will be at $V_{ref}/2$ which is 0.5 V. This poses a difficulty in the design of the reference switch. Normally, if a PFET is used as a reference switch to pull up the node V_{comp} to V_{ref}, its bulk terminal must be connected to the highest voltage in the circuit to prevent the body diodes from conducting. However, in the situations discussed earlier, it is unclear what the highest voltage in the circuit is. For the case when the ADC samples the input signal $V_{in} = 0$ V, V_{comp} can reach 1.5 V during the first approximation by the SAR logic. Thus, if we use a PFET whose bulk terminal is connected to the ADC's power supply of 1 V as a reference switch, the 0.5 V difference between the source and the bulk terminals of the PFET can forward bias its source-to-bulk diode. As a result, the charge stored on the capacitor DAC array can get discharged through this forward-biased diode, resulting in incorrect decisions during the bit-cycling periods. To ensure that the reference switch remains open and the total charge on the capacitor DAC array is preserved throughout the bit cycling periods, we utilize a bootstrapping technique in which the bulk's and the gate's voltages of a

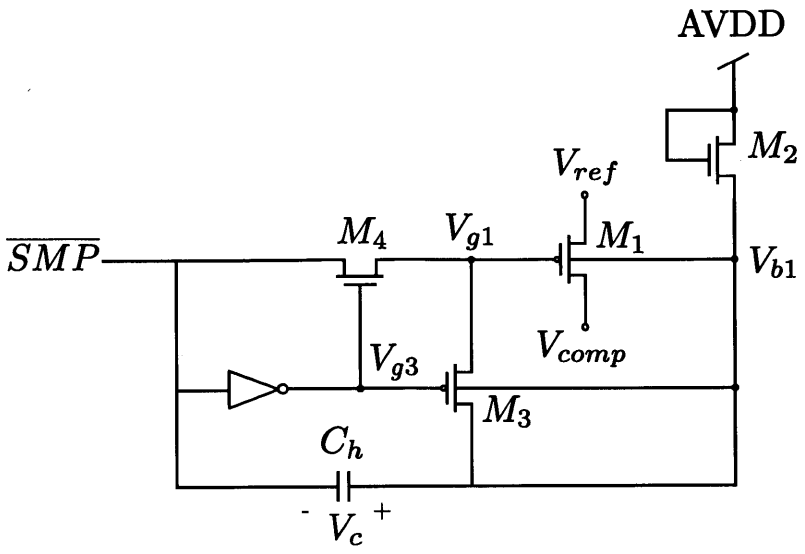


Figure 3-16: Schematic of the bootstrapped reference switch.

PFET switch are bootstrapped to higher than the ADC's supply voltage such that even if V_{comp} reaches 1.5 V, the reference switch and the body diode remain strongly off, preserving the total charge on the capacitor DAC array throughout the whole conversion period.

The bootstrapped reference switch is shown in Fig. 3-16 [57]. The PFET M_1 is the core of the reference switch. The operation of the switch can be described as follows: Let's assume that, originally, the voltage across the capacitor C_h, V_c, is zero. During the sampling duration, the signal $\overline{SMP}$ goes low making V_{g3} go high, turning on M_4 and turning off M_3. The gate voltage of M_1, V_{g1}, is thus pulled down to ground by M_4, turning on the core switch M_1. During this sampling duration, the bottom plate of C_h is pulled down to ground and the bulk voltage V_{b1} is originally pulled down to zero by the uncharged capacitor C_h, turning on the transistor M_2. As a result, M_2 starts sourcing current to charge the capacitor C_h, raising M_1's bulk voltage, V_{b1}. The voltage V_{b1} rises until M_2 enters the cutoff region, which is approximately when V_{b1} reaches AVDD $- V_{tn2}$, where V_{tn2} is the threshold voltage of M_2. At this point, the voltage across the capacitor, V_c, is approximately AVDD $- V_{tn2}$. Once the sampling duration ends, $\overline{SMP}$ steps up to AVDD. The voltage V_{g3} is pulled low by the inverter, turning on M_3, while turning off M_4. At this point, V_{g1} is shorted to V_{b1} by the switch

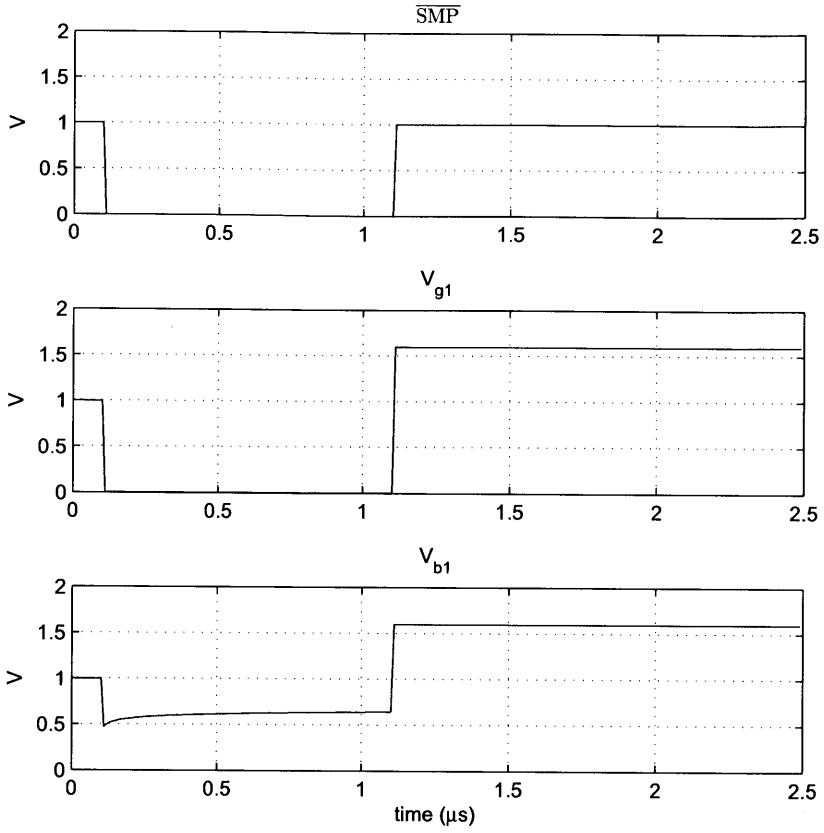


Figure 3-17: Simulation result showing the node voltages of the reference switch in Fig 3-16.

M_3. Since the voltage across C_h is AVDD $- V_{tn2}$, the voltage at nodes V_{b1} and V_{g1} goes to 2AVDD $- V_{tn2}$ when $\overline{SMP}$ steps up to AVDD. In the process technology that this SAR ADC is implemented, the threshold voltage is approximately $V_{tn2} = 500$ mV. Therefore, the gate and the bulk voltages of the core switch M_1 are bootstrapped to approximately 1.5 V. In the worst case when $V_{in} = 0$ V where V_{comp} can reach 1.5 V, V_{g1} and V_{b1} are high enough to ensure that the source-to-bulk diode of M_1 remain strongly off during the bit cycling periods.

Fig. 3-17 shows the simulation results of the waveforms at the nodes $\overline{SMP}$, V_{g1}, and V_{b1}. The sampling duration occurs between $t = 0.1$ μs to $t = 1.1$ μs. During this sampling duration, the voltage at V_{b1} is charged up to around 640 mV, which is

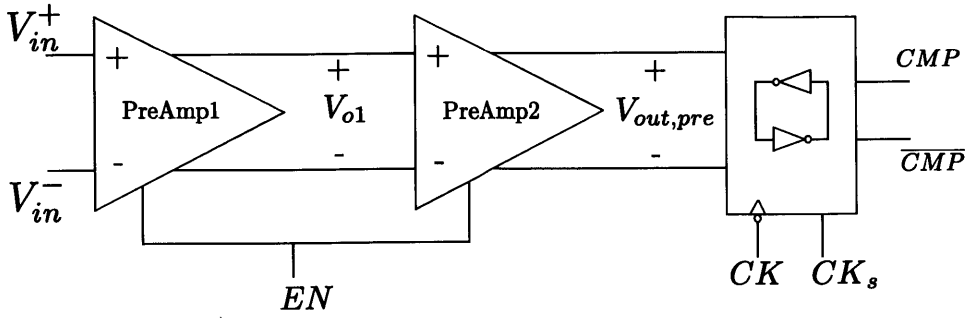


Figure 3-18: Schematic of the comparator.

even higher than AVDD $- V_{tn}$. This is because the capacitor C_h is charged by the subhtreshold current of M_2 even after M_2 enters the cutoff region. As a result, when $\overline{SMP}$ goes high, V_{b1} and V_{g1} jump to higher than 1.5 V, which ensure that the switch M_1 remain strongly off throughout the bit cycling periods.

Comparator

The schematic of the comparator is shown in Fig. 3-18. The comparator consists of a preamplification stage followed by a latch. The preamplification stage consists of two amplifiers connected in a cascade manner. Preamplification is required to amplify the small voltage difference between V_{comp} and V_{ref} to overcome the relatively large input-referred offset voltage of the latch, and also to prevent the kickback from the latch which may corrupt the sampled voltage on the capacitor DAC array. To improve the sensitivity of the preamplifiers, the power supply pins of the sensitive preamplifiers and the noisy latch are decoupled. The sensitive preamplifiers operate from an analog supply voltage, AVDD, while the noisy latch operates from a digital supply voltage, DVDD. The nominal value of both AVDD and DVDD is 1 V.

The schematic of the preamplifier is shown in Fig. 3-19. The preamplifier utilizes an NMOS input differential pair with PMOS transistors operating in the triode region as resistive loads. The transistor M_{b1} sets the bias current which determines the gain, speed, input-referred noise, and power consumption of the preamplifier. The transistor M_{b2} acts as a duty cycling switch which can shut down the bias current in the preamplfier when it is not needed. For a moderate-precision ADC such as this

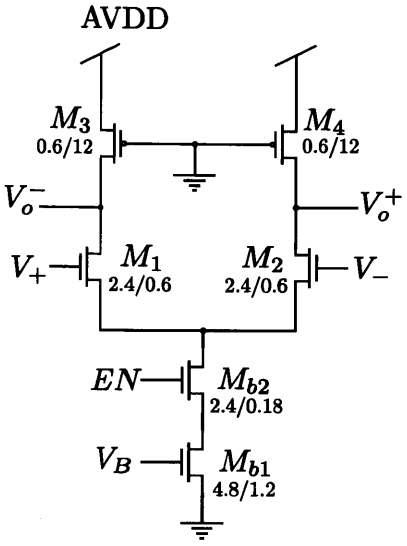


Figure 3-19: Schematic of the preamplifier.

one, the intrinsic noise of the preamplification stage is usually not the performance limiting factor. This statement will be verified later in this section. To determine a suitable bias current level for the preamplifiers, first let's consider the requirements for the gain and speed of the preamplifiers. The total gain of the preamplification stage should be large enough such that the comparator is able to resolve the input voltage as small as $V_{LSB}/2$ into a correct logical decision. Thus, it requires that the preamplification stage must be able to amplify the input signal of value $V_{LSB}/2$ to larger than the input-referred offset $V_{off,L}$ of the latch. If we denote the gains of the first and the second preamplifiers as A_{pre1} and A_{pre2} respectively, the gain requirement of the preamplification stage can be stated simply as

$$A_{pre1} \cdot A_{pre2} \cdot V_{LSB}/2 > V_{off,L}. \tag{3.27}$$

Since the full-scale voltage of the ADC is $V_{ref} = 1$ V, and for an 8-bit ADC, $V_{LSB} = V_{ref}/2^8$, (3.27) can be re-written as

$$A_{pre1} \cdot A_{pre2} > \frac{2^9 \cdot V_{off,L}}{V_{ref}}. \tag{3.28}$$

To improve the speed of the preamplification stage, the input transistors of the latch

are made with near minimum-sized transistors such that they do not load the output nodes of the second preamplifiers. With small input transistors, the input-referred offset voltage of the latch can be as high as 50 mV. Thus, the gain requirement for the preamplification stage can be stated as $A_{pre1} \cdot A_{pre2} > 25$.

Besides providing enough gain, the preamplifiers must operate fast enough to ensure that the comparator can make a decision within the required time. Each bit cycling period lasts one period of CK, with $T_{CK} = 400$ ns. Preamplification occurs during the first half of the CK period (when CK goes high), while the latching occurs during the second half (once CK goes low). Thus, the output $V_{out,pre}$ of the preamplification stage in Fig. 3-18 must settle close to its final value within $T_{CK}/2 = 200$ ns. The preamplification stage behaves as a second-order system with two dominant poles close to each other in frequency. One pole is associated with the node V_{o1} of the first preamplifier while another pole is associated with the node $V_{out,pre}$. Nevertheless, for simplicity in analyzing the settling time requirement, let's assume that the preamplification stage behaves approximately as a first-order linear amplifier. The time required for the output $V_{out,pre}$ to settle to 90 % of its final value when excited with a step input is $t_{90\%} = \tau_{pre} \cdot \ln(10)$ where τ_{pre} is the effective time constant of the overall preamplification stage. The time constant τ_{pre} can be approximated using the Open-Circuit time constant analysis [21], which can be written as $\tau_{pre} = R_{o1}C_{o1} + R_{o2}C_{o2}$ where R_{oi} and C_{oi} are the output resistance and the output capacitance of the i^{th} preamplifier respectively. Therefore, the speed requirement of the preamplification stage can be written as

$$\tau_{pre} < \frac{T_{CK}}{2 \cdot \ln(10)} = 87 \text{ ns.} \tag{3.29}$$

The -3-dB cutoff frequency of a first-order linear amplifier can be related to its time constant by the relationship $f_{-3dB} = 1/2\pi\tau_{pre}$. Thus, the bandwidth requirement of the preamplification stage for this ADC is $f_{-3dB} > 1.8$ MHz.

From the schematic of Fig. 3-19, the differential gain of the preamplifier can be

approximated by

$$\begin{aligned} A_{pre} &= g_{m1}/(g_{ds1} + g_{ds3}) \\ &\approx g_{m1}/g_{ds3}. \end{aligned} \tag{3.30}$$

where g_{m1}, g_{ds1}, and g_{ds3} are the transconductance of M_1, the drain-to-source admittance of M_1, and the drain-to-source admittance of M_3 respectively. In (3.30), we made an assumption that $g_{ds3} \gg g_{ds1}$ because M_3 and M_4 operate in the triode region while M_1 and M_2 operate in the saturation region. Using a long-channel model of a MOSFET, we can write g_{ds3} as

$$g_{ds3} = \mu_p C_{ox} \frac{W_3}{L_3} \left(V_{SG3} - |V_{T3}|\right) \tag{3.31}$$

where μ_p is the hole mobility, V_{SG3} is the source-to-gate voltage of M_3, and V_{T3} is the threshold voltage of M_3. Since $V_{SG3} = \text{AVDD}$, the expression in (3.30) can be written as

$$A_{pre} = \frac{g_{m1}}{\mu_p C_{ox} \frac{W_3}{L_3} \left(\text{AVDD} - |V_{T3}|\right)}. \tag{3.32}$$

The input-referred thermal noise of each preamplifier $\overline{v_{n,pre}^2}$ can be written as

$$\overline{v_{n,pre}^2} = \frac{16}{3} \frac{kT}{g_{m1}} + \frac{1}{g_{m1}^2} 8kT\mu_p C_{ox} \frac{W_3}{L_3} \left(\text{AVDD} - |V_{T3}|\right). \tag{3.33}$$

Due to a high gain of the preamplification stage, its noise dominates the total noise of the comparator. Thus, to minimize the input-referred noise of the comparator, we have to minimize the noise contributions from the preamplifiers. From (3.32) and (3.33), we can both minimize $\overline{v_{n,pre}^2}$ and at the same time maximize A_{pre} by maximizing g_{m1} and minimizing W_3/L_3. To maximize g_{m1}, M_1 and M_2 are sized with large W/L ratio such that they operate in subthreshold, while M_3 and M_4 are sized with small W/L ratio such that they operate deep in strong inversion and their noise contributions are minimized. However, it must be cautioned that making W_3/L_3 small can degrade the speed of the comparator due to the increase in the output resistances

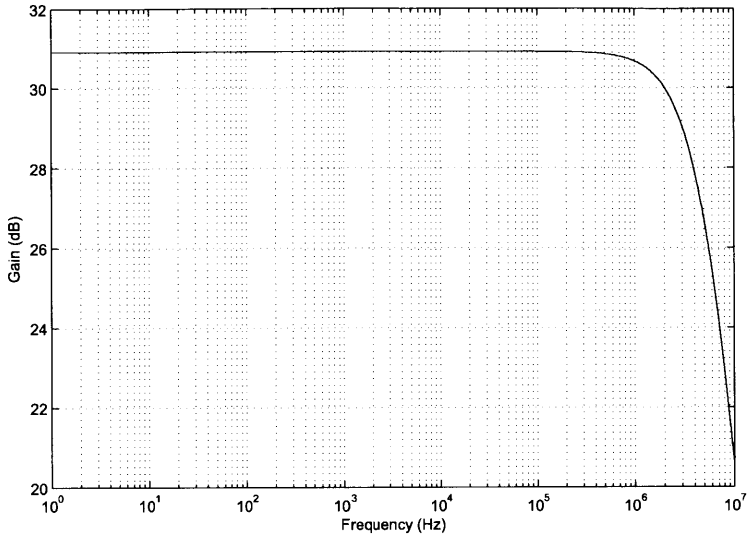


Figure 3-20: Magnitude response of the preamplification stage. The low-frequency gain is 30.9 dB (35) and $f_{-3\text{dB}} = 4$ MHz

of the preamplifiers. Equations (3.29), (3.32), and (3.33) can be used along with circuit simulations to design the preamplification stage that meets the speed, gain, and noise requirements of our application.

The sizes of the transistors in the preamplifier are indicated in Fig. 3-19. When the preamplifier are biased with the current in M_{b1} of approximately 600 nA, the small signal parameters from a DC simulation are $g_{m1} = 6.9\mu$A/V, $g_{ds1} = 154$ nA/V, and $g_{ds3} = 986$ nA/V. According to (3.30), the DC gain per each amplifier is calculated to be 6. Figure 3.3.4 shows the magnitude response from an AC simulation of the preamplification stage. The low-frequency gain of the preamplification stage is 35.1 (30.9 dB), while the -3-dB cutoff frequency is approximately 4 MHz. Figure 3-21 shows the simulated output noise density at the node $V_{out,pre}$ of Fig. 3-19. This noise density results in a total noise of approximately 7 mV_{rms} at $V_{out,pre}$. The input-referred noise of the preamplification stage is thus (7 mV_{rms})/35 = 200 μV_{rms}. Please note that the quantization noise of our 8-bit SAR ADC is approximately 1.13 mV_{rms}. Thus the intrinsic noise of the preamplification stage, and thus the comparator, accounts for only 17 % of the ADC's quantization noise and can safely be ignored.

According to the timing diagram in Fig 3-14(b), the sampling period of the ADC is

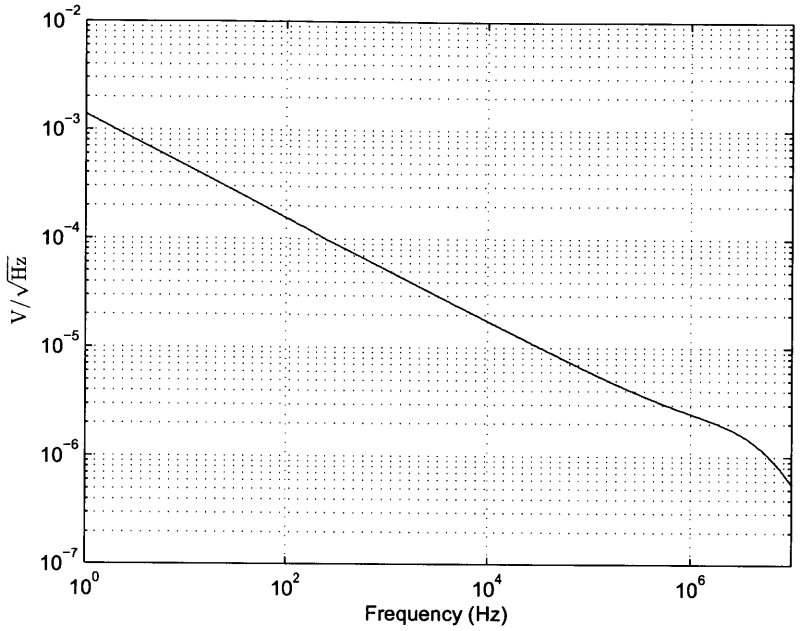


Figure 3-21: Output noise density at $V_{out,pre}$.

$20T_{CK}$, while the conversion process in which the comparator is needed is only about half that long. To minimize power consumption in the preamplifiers, their supply currents are shut down when they are not needed. The transistor M_{b2} in Fig. 3-19 whose gate is connected to the enable signal EN is used as a duty cycling switch for each preamplifier. When EN is high, both preamplifiers are turned on. Note that the time at which the preamplifiers are turned on is 4.1 μs, which is about 50 % of the 8 μs sampling period. As a result, a power saving by a factor of 2 can be achieved compared to when both preamplifiers are on at all time.

The schematic of the latch is shown in Fig. 3-22. As mentioned earlier, when CK is high, the voltage difference between V_{ref} and V_{comp} is amplified by the preamplification stage to overcome the input-referred offset of the latch. During this time, the latch is non-operational. The transistors M_{r1}, M_{r2} and M_{r3} are on, pinning the drains of M_5 and M_6 to the supply voltage DVDD while shorting the drains of M_1 and M_2 together. At this time, M_B is turned off such that no static current is wasted in the latch. The latch is now in the balance state. Immediately when CK goes low, M_B is turned on while M_{r1}, M_{r2} and M_{r3} are turned off. The transistors M_3, M_5 and M_4,

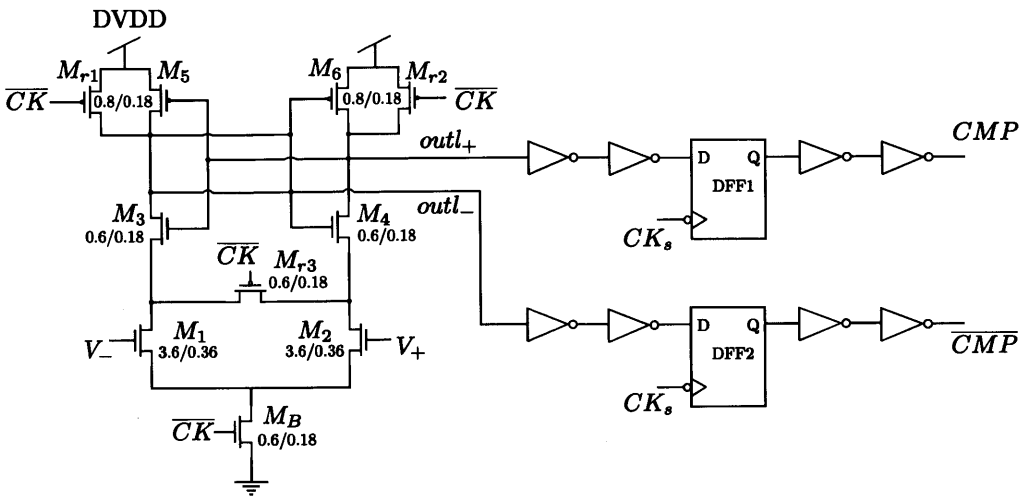


Figure 3-22: Schematic of the latch.

M_6 behave as cross-coupled inverters forming a positive feedback loop. The voltages on the nodes $outl_+$ and $outl_-$ are discharged by M_1 and M_2 at an unequal rate which is determined by the differential voltage $V_+ - V_- = V_{out,pre}$, thus creating an unbalance in the cross-coupled inverters. Once the difference between $outl_+$ and $outl_-$ is large enough, the positive feedback action kicks in. The latch starts to regenerate, latching the voltages $outl_+$ and $outl_-$ to the opposite supply rails. The SAR logic makes the decision on the positive edge of CK which is exactly the moment when the nodes $outl_+$ and $outl_-$ are reset to DVDD. To ensure that the comparator's outputs CMP and $\overline{CMP}$ are always constant at the positive edge of CK, which is the moment of decision by the SAR logic, we use the output registers DFF1 and DFF2 to sample the latch's outputs at the negative edge of an auxiliary clock signal CK_s as shown in Fig. 3-14(b). As a result, the comparator's outputs CMP and $\overline{CMP}$ will have been stable for about 100 ns before each positive edge of CK.

Custom SAR logic and Switch Network

The schematic of the SAR logic is shown in Fig. 3-23. It consists of a chain of shift registers (labeled SR) and eight switch drive registers (labeled SDR). The shift register chain behaves as a state machine that controls the order of operation of the ADC through the index signals L_7 ($\overline{L_7}$) to L_0 ($\overline{L_0}$). The index signals L_7-L_0 act as

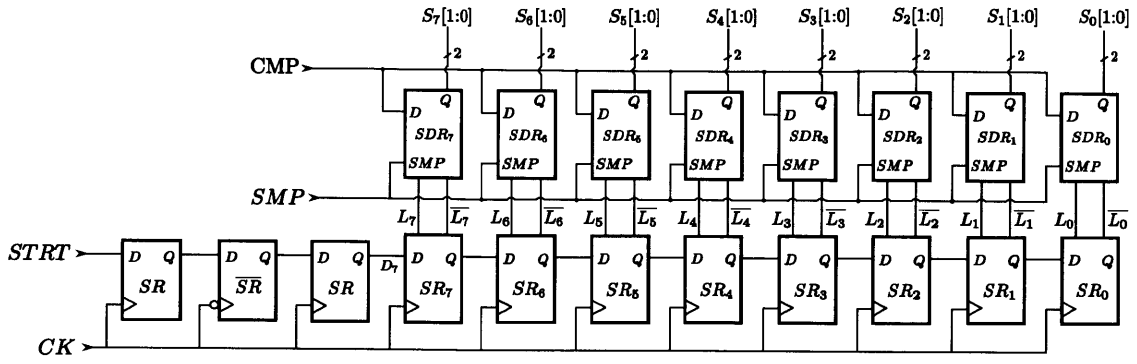


Figure 3-23: Schematic of the SAR logic.

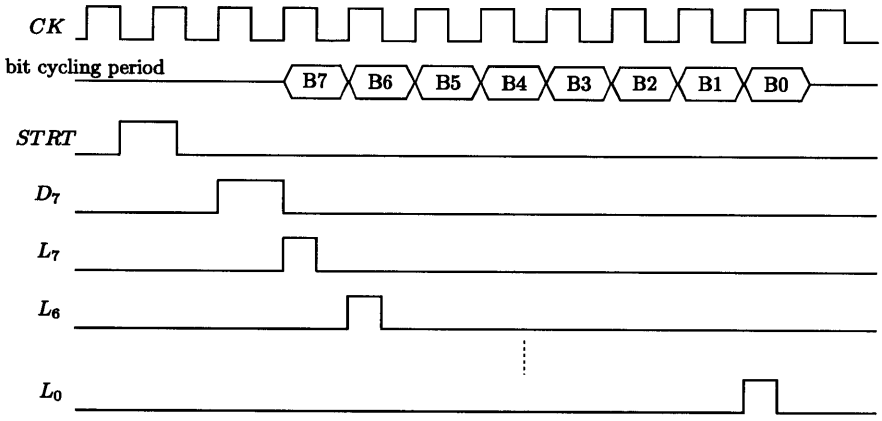


Figure 3-24: Timing diagram of the SAR logic.

enable signals for SDR_7-SDR_0; when L_i ($i \in \{0, ..., 7\}$) is asserted, the register SDR_i asserts the output $S_i[1:0]$ to configure the switches in bit cell i of Fig. 3-14(a). The timing diagram of the SAR logic is given in Fig. 3-24. At the start of the conversion process, a $STRT$ pulse is asserted at the input of the leftmost shift register SR to initiate the conversion. The pulse is clocked through the shift register chain at every positive edge of the clock signal CK. During the bit cycling period Bi, the pulse reaches the shift register SR_i, and the index signal L_i is asserted for half the period of CK. The signal L_i then instructs the corresponding switch drive register SDR_i to determine the states of the switch drive signals $S_i[1:0]$, based on the value of the comparator's output signal CMP from the previous decision.

To minimize power consumption of the SAR logic, the shift registers and switch drive registers are custom designed using dynamic logic techniques [47] to minimize

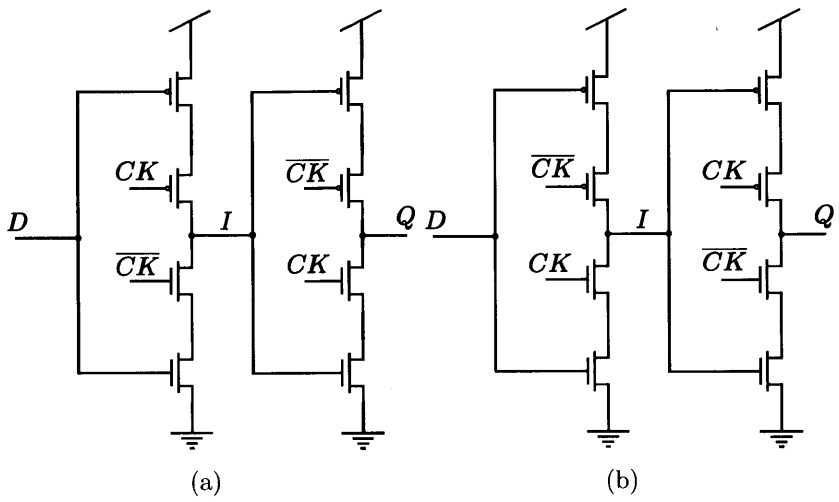


Figure 3-25: (a) Schematic of SR (b) Schematic of $\overline{SR}$

the number of internal capacitances that need to be switched. The schematics of SR and $\overline{SR}$ based on the C^{2}MOS registers [47] are shown in Fig. 3-25(a) and 3-25(b) respectively. The register SR is a positive-edge triggered version while the register $\overline{SR}$ is a negative-edge triggered version. These registers are dynamic in nature because they utilize internal capacitances as the storage mechanism instead of the positive feedback commonly employed in static logic circuits. For the register SR in Fig. 3-25(a), when CK goes low, the internal node I receives the value of $\overline{D}$ while node Q is at high impedance. Once CK goes high, the internal node I becomes high impedance and the value of $\overline{D}$ right before the positive edge of CK is stored in the internal capacitance at the node I. The output Q becomes low impedance and takes on the value of $\overline{I}$, which is equal to the value of D just right before the positive edge of CK. The operation of $\overline{SR}$ is the same as that of SR except that the phase of CK is reversed. The output Q of $\overline{SR}$ takes on the value of the input D just right before the negative edge of CK.

The schematic of the shift register SR_i is shown in Fig. 3-26. It is also based on the C^{2}MOS topology, but with extra transistors to ensure that the internal nodes are properly reset or precharged to appropriate values such that the output Q and L_i are reset to ground after the $STRT$ signal is asserted ($\overline{STRT}$ is low). This requirement

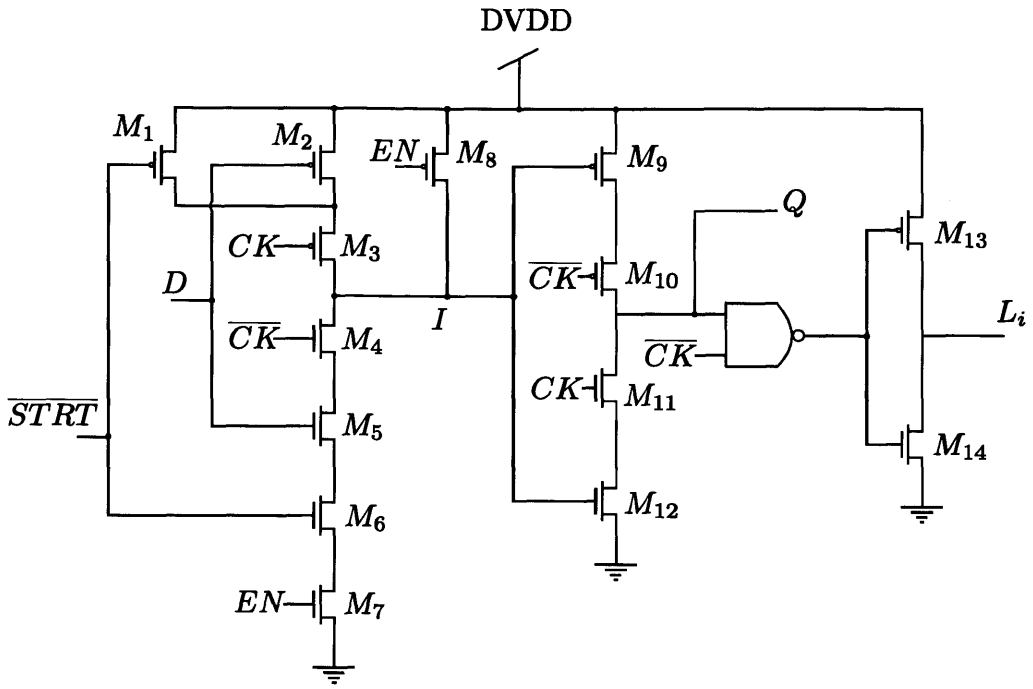


Figure 3-26: Schematic of the shift register SR_i.

is very important because the switch drive register SDR_i that is driven by L_i is designed with dynamic logic techniques; if L_i is incorrectly asserted at the wrong time, the switch drive signals $S_i[1:0]$ may be accidentally switched to the wrong values, and cannot be restored to the correct values until the next conversion period. This scenario can result in an error of the conversion and must be avoided. To prevent such error from happening, we make sure that within one period of CK after $STRT$ is asserted, the output Q and L_i of all the shift registers SR_i are reset to 0. During the conversion period, the EN signal is high, turning on M_7 and turning off M_8. When $STRT$ is high ($\overline{STRT}$ is low) and when CK goes low, the transistors M_1 and M_3 are on while the transistor M_6 is off. As a result, the intermediate node I is precharged through M_1 and M_3 to DVDD, regardless of the value of the input D. When CK goes high, the tri-state inverter formed by M_9-M_{12} becomes transparent, driving the output Q to a logic 0. As a result, the output L_i also becomes 0 regardless of the logic level of CK. After the $STRT$ signal goes low ($\overline{STRT}$ is high), M_1 is turned off while M_6 is turned on. The register SR_i now behave as a regular C^2MOS shift register with the input D and the output Q. The NAND gate and the inverter formed

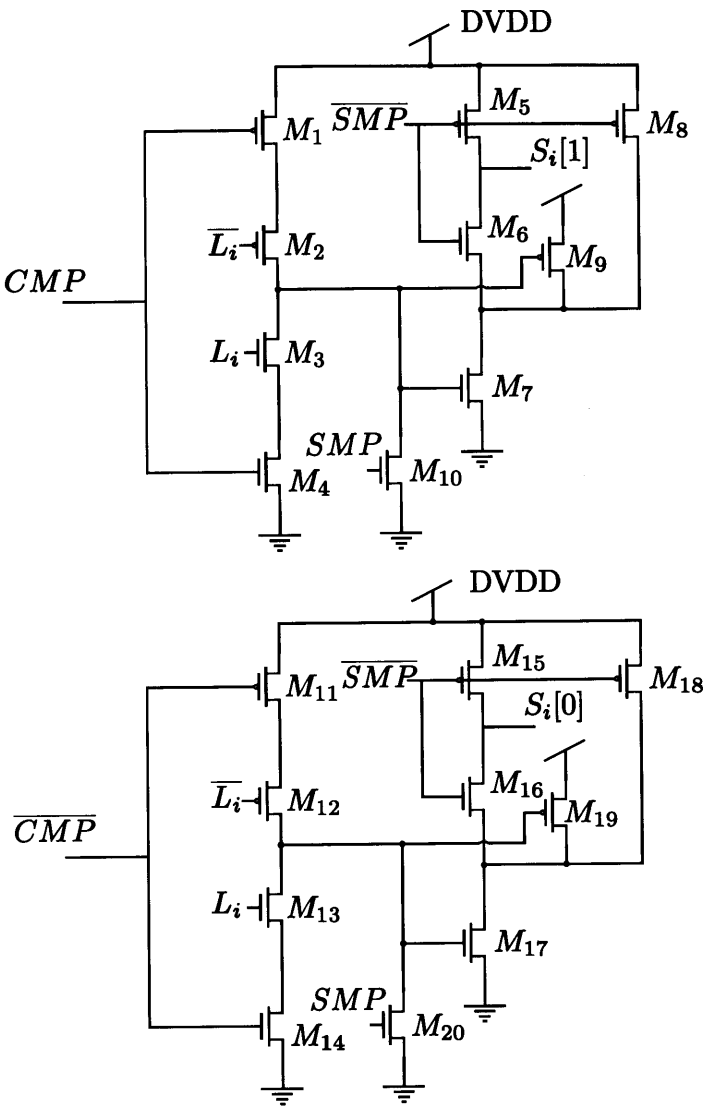


Figure 3-27: Schematic of the switch drive register SDR_i.

by M_{13} and M_{14} are included for creating L_i such that L_i goes high only when both Q and CK are high. This feature of L_i is to guarantee the correct operation of the switch drive register SDR_i that is driven by L_i, and will be explained later in this section. The schematic of the switch drive register SDR_i is shown in Fig. 3-27. It consists of two identical dynamic registers for separately controlling $S_i[1]$ and $S_i[0]$. During the sampling duration when SMP is high ($\overline{SMP}$ is low), the transistors M_5, M_8, M_{15}, and M_{18} are turned on, while M_6, M_7, M_{16} and M_{17} are turned off. As a result, $S_i[1]$ and $S_i[0]$ are precharged to DVDD through M_5 and M_{15} respectively.

Since the source nodes of M_6 and M_{16} are at high impedance and sit at undefined voltages when SMP is high, we add M_8, M_9 and M_{18}, M_{19} to precharge these nodes to DVDD. This is to prevent the charge sharing effect when SMP goes low in which the charge at the nodes $S_i[1]$ and $S_i[0]$ will be distributed to the parasitic capacitances at the source nodes of M_6 and M_{16}, and thus lowering the voltages at $S_i[1]$ and $S_i[0]$. The original state of the switch drive signals in which $S_i[1]$ and $S_i[0]$ are both at logic 1 corresponds to the switch configuration where the bottom plate of the left capacitor in bit cell i is connected V_{ref} and the bottom plate of the right capacitor in bit cell i is connected to ground. During the decision process when L_i goes high, only one of either $S_i[1]$ or $S_i[0]$ will be discharged to ground, thus only one of the capacitors in bit cell i will change its configuration. If $S_i[1]$ is discharged to ground, the left capacitor in the bit cell i, which is originally connected to V_{ref}, will be switched to ground. Similarly, if $S_i[0]$ is discharged to ground, the right capacitor in bit cell i, which is originally connected to ground, will be switched to V_{ref}. Note that when L_i goes high, the outputs of the comparator CMP and $\overline{CMP}$ are already stable (with the helps of DFF1 and DFF2 in Fig. 3-22). Consider the top register of Fig. 3-27. If CMP is low ($\overline{CMP}$ is high) when L_i goes high, the gate voltage of M_7 will be pulled up to DVDD through M_1 and M_2, turning on M_7 and turning off M_9. Since M_6 is already on once SMP goes low, the voltage at $S_i[1]$ will be discharged to ground. However, for the bottom register of Fig. 3-27, because $\overline{CMP}$ is high, the gate voltage of M_{17} is pinned to ground through M_{13} and M_{14}, causing M_{17} to remain in the off state. The voltage at node $S_i[0]$ thus remains at logic 1 as desired. On the contrary, if CMP is low when L_i goes high, $S_i[0]$ will be discharged to ground while $S_i[1]$ will remain at logic 1.

The schematics of the switch network is shown in Fig. 3-28. The nodes B_{iM} and B_{iL} drive the bottom plates of the left and right capacitors in bit cell i respectively. During the sampling duration when SMP goes high, the transmission gate switches, formed by M_2, M_3 and M_6, M_7, connect both B_{iM} and B_{iL} to the input V_{in}. Please recall from the description of SDR_i that both $S_i[1]$ and $S_i[0]$ are already precharged to logic 1 during the sampling duration. As a result, when SMP goes low, B_{iM}

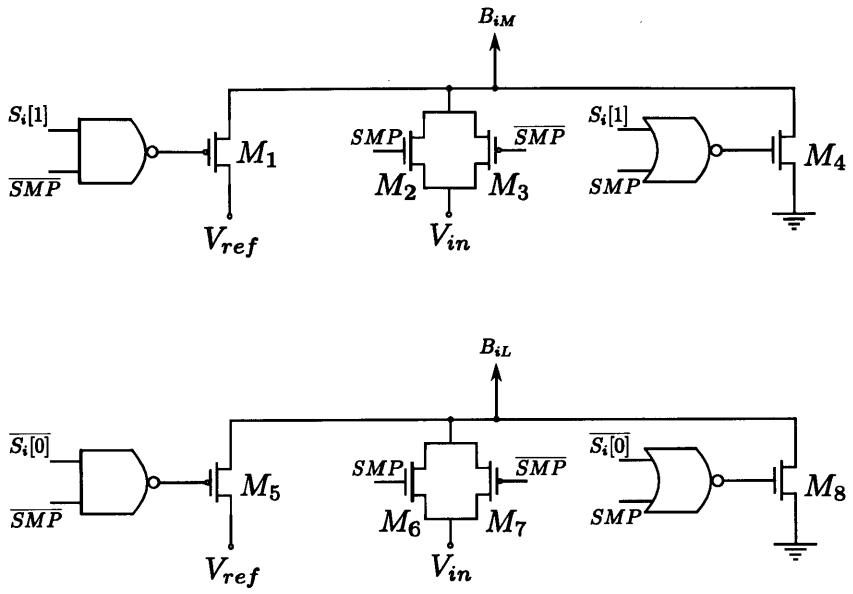


Figure 3-28: Schematic of the switch network.

is connected to V_{ref} through the transistor M_1, while B_{iL} is connected to ground through the transistors M_8. During the bit cycling period Bi when L_i goes high (SMP is already low), one and only one of the switch drive signals $S_i[1]$ or $S_i[0]$ is discharged to a logic 0. If $S_i[1]$ is 0, the transistor M_1 is turned off while M_4 is turned on, and thus B_{iM} is switched from V_{ref} to ground. On the other hand, if $S_i[0]$ is discharged to a logic 0, M_8 is turned off while M_5 is turned on, and thus B_{iL} is switched from ground to V_{ref}. As a result, the reconfiguration of the capacitors in bit cell i according to the flow graph in Fig. 3-15 is accomplished.

3.4 Analog Multiplexer

The outputs from four amplifiers are multiplexed through an analog multiplexer onto the ADC in the recording module. The analog multiplexer needs to be strong enough to drive the input capacitance of the ADC. Furthermore, the input voltage range of the ADC is from 0 to $V_{ref} = 1$ V while the DC level of the amplifier's output is $V_{mid} =$ 0.9 V. To utilize the whole input range of the ADC, the analog multiplexer must therefore also perform a DC level shifting function such that the input to the ADC is

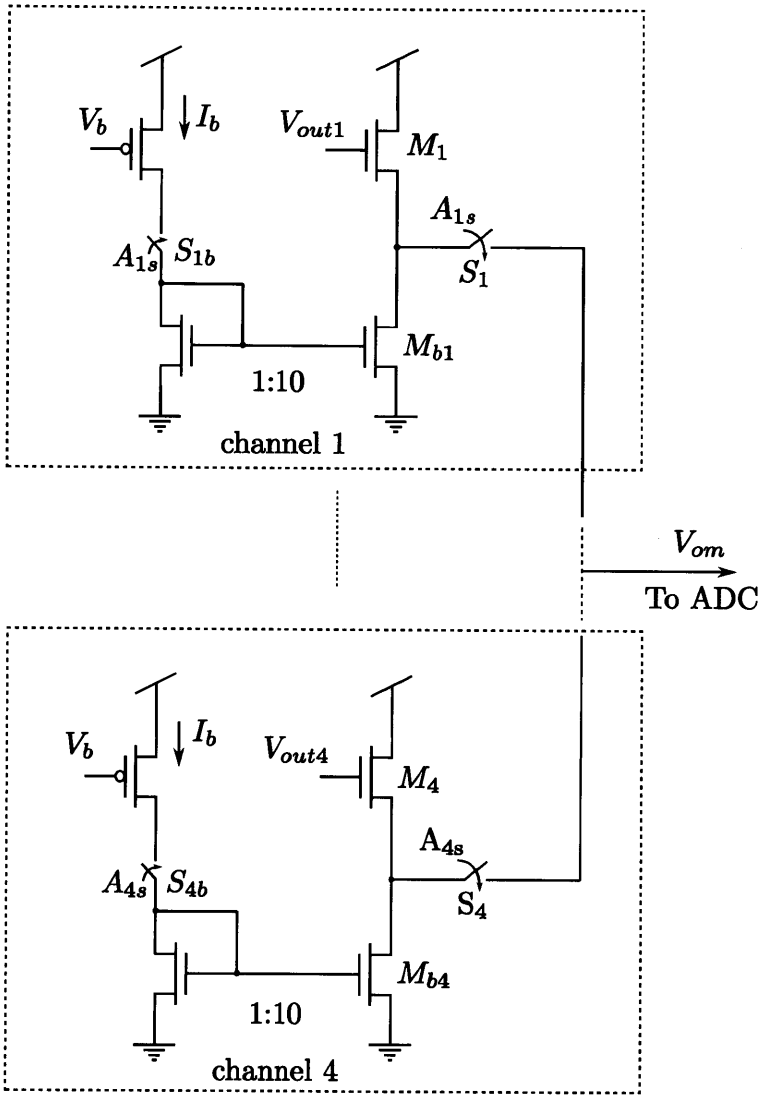


Figure 3-29: Schematic of the analog multiplexer.

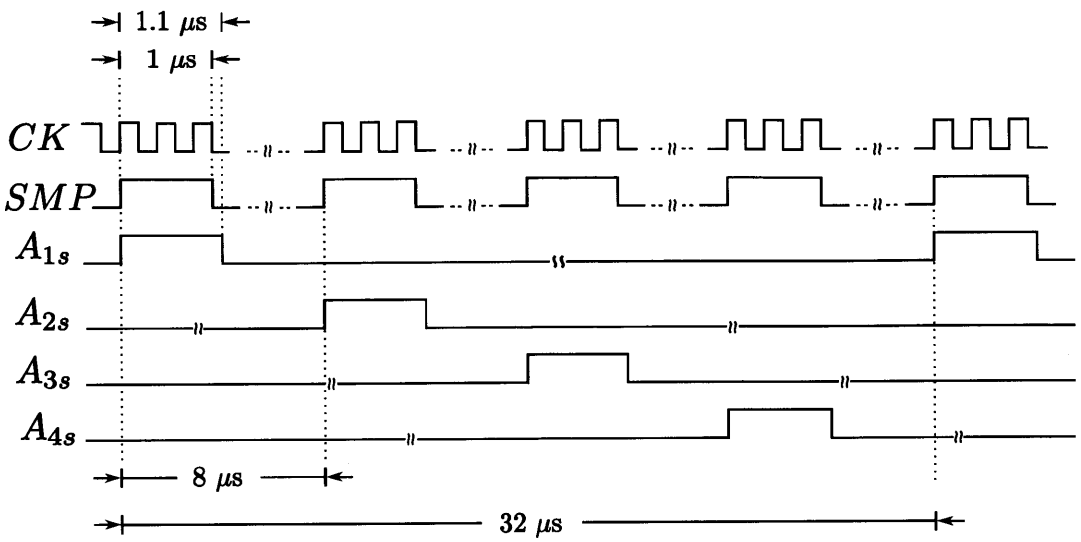


Figure 3-30: Timing diagram of the analog multiplexer's control signals.

centered near the midpoint of the ADC's input range (0.5 V). Figure 3-29 shows the schematic of the analog multiplexer. The core of the analog multiplexer consists of four source-follower drivers which are formed by the transistors M_1, M_{b1},...,M_4, M_{b4}. The source-follower drivers buffer the amplifiers' outputs V_{out1}-V_{out4} to provide low output impedance necessary for driving the input capacitance of the ADC and for also performing DC level shifting function that centers the ADC's input near 0.5 V. Multiplexing the amplifiers' outputs onto the ADC is achieved through the switches S_1-S_4 which are controlled by the control signals A_{1s}-A_{4s} respectively. When A_{is} ($i \in \{1, ..., 4\}$) is high, the switch S_i is closed and V_{outi} is buffered and multiplexed into the input of the ADC. The timing diagram of the control signals $A_{1s}, ..., A_{4s}$ is shown in Fig. 3-30. Note that only one of the switches S_1-S_4 is closed at a given time. The duration at which each switch is closed is 1.1 μs which suffices to span the ADC's sampling duration (when SMP is high for 1 μs). As a result, the ADC's input is driven by only one low-impedance source at the sampling instant of the ADC (at the negative edge of SMP).

In order to drive the ADC's input capacitance, the source-follower driver must achieve a low output impedance, thus requiring a high bias current. Fortunately, each source-follower driver only needs to be active when it is driving the input capacitance

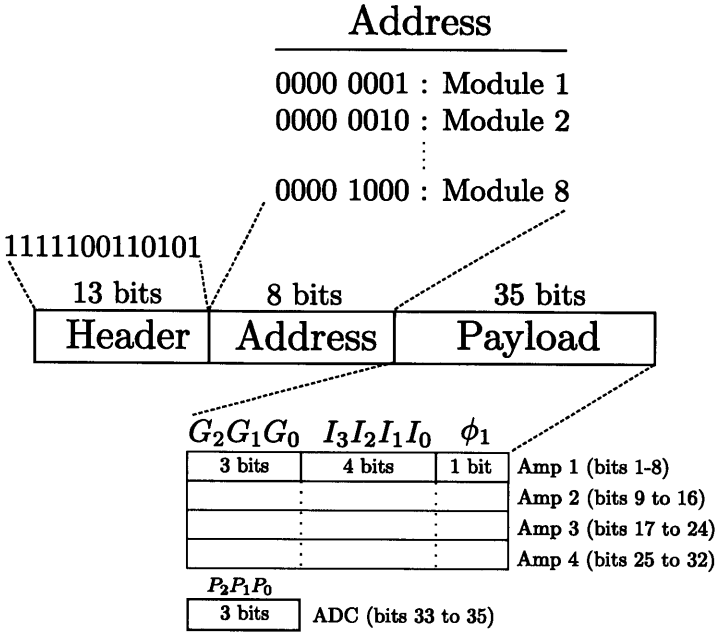


Figure 3-31: Format of the programming packet for configuring each neural recording module.

of the ADC, which is only 1.1 μs for the whole 32 μs duration of the sampling period. By turning on each source-follower only when it is driving the ADC's input, we can achieve the power saving by almost a factor of 32 compared to the case when all the drivers are constantly powered. We therefore duty cycle the bias current in the i^{th} source-follower driver by its corresponding control signal A_{is}. When A_{is} is high, the switch S_{ib} is closed. The current I_b is mirrored with a 10x gain to the i^{th} source follower driver. The ADC's sampling period of 1 μs allows the bias current in the source-follower driver to settle before the sampling instant takes place at the negative edge of SMP. When A_{is} goes low, the switch S_{ib} is opened, turning off the bias current in the the source follower driver to save power.

3.5 Serial Programming Interface

Configuring the 32-channel neural recording IC is achieved on a module-by-module basis. As shown in Fig. 3-2, each recording module contains a dedicated serial programming interface unit. In each recording channel, the bias current of the front-end

amplifier, the gain of the programmable-gain amplifier, and the recording setting are configured with 4-bit, 3-bit, and 1-bit control inputs respectively. For the purpose of calibration against process variations, the bias currents of the preamplifiers in the ADC can also be controlled by a 3-bit control input. To configure the recording channel's setting and the ADC's preamplifiers in each recording module, the user must provide a 56-bit programming data stream in a recognizable format as shown in Fig. 3-31. The 56-bit programming data stream consists of a 13-bit header field, an 8-bit module address, and a 35-bit payload field. The header field is a fixed sequence "1111100110101". The module address is an 8-bit binary value that specifies the address (from 1 to 8) of the recording module to be programmed. Once the header field and the module address are recognized by the internal logic, the 35 payload bits are clocked into storage registers that provide the control inputs to the recording channels and the ADC's preamplifiers. The programmed bits are maintained as long as the power is supplied to the 32-channel neural recording IC or until the user reprograms the recording module with different payload.

The circuit architecture of the serial programming interface unit is shown in Fig. 3-32. The programming data (prog data) is shifted into a 56-bit shift register on the positive edge of the programming clock (prog clock). Two digital comparators are used to detect whether the recognition sequence and the module address match the specified values. If both match (recog_match and address_match signals go high), the load signal, load_prog, is asserted, instructing the storage registers to load the programming bits from the payload field to their outputs. In addition, the signal load_prog is used to clear the 56-bit shift register to guarantee that, once the recognition sequence and the module address match, the payload field is cleared such that it cannot trigger another loading.

3.6 Digital Control Unit

In this section, we present the design of the Digital Control Unit. The Digital Control unit serves two purposes. First, it creates all the control signals for the ADCs and

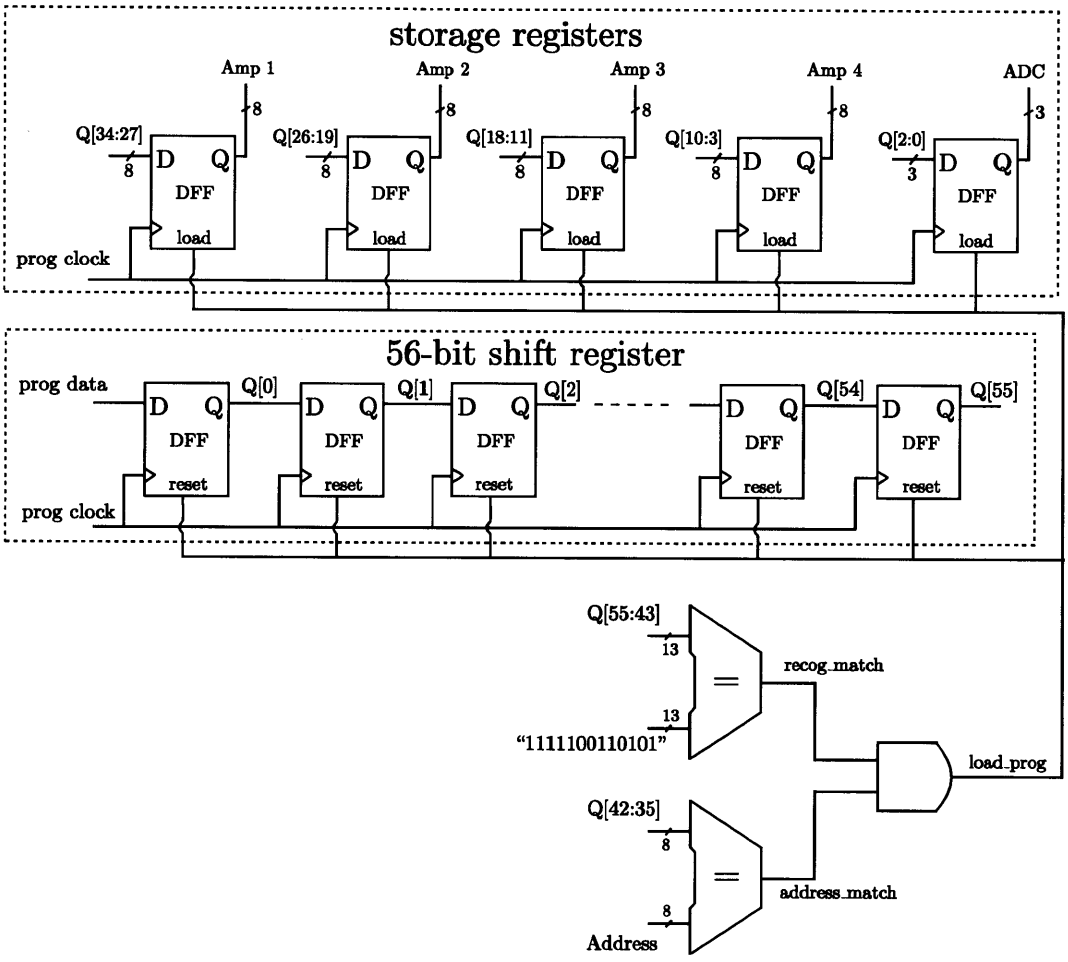


Figure 3-32: Circuit architecture of the serial programming interface unit.

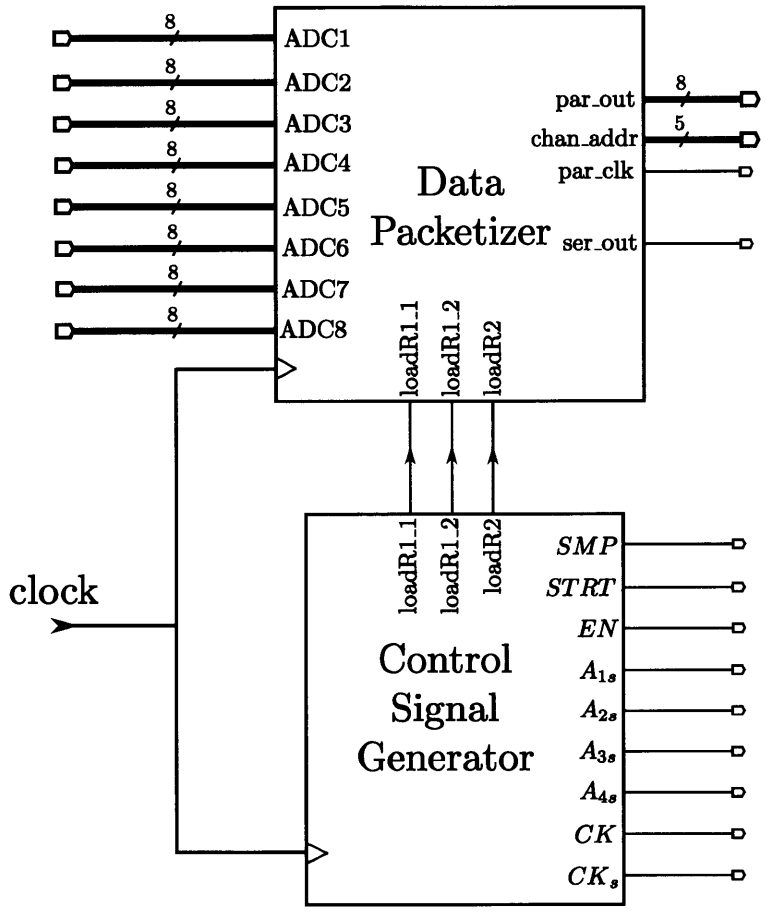


Figure 3-33: Block diagram of the Digital Control Unit.

analog multiplexers on the 32-channel neural recording IC. Second, it organizes the outputs from the eight ADCs, before sending these data off-chip to the on-board FPGA for further processing. We decide not to implement a specific digital signal processing unit on chip to allow for design flexibility obtained from using an on-board FPGA. The block diagram of the Digital Control Unit is shown in Fig. 3-33. It consists of two separate units: the Control Signal Generator and the Data Packetizer. The Control Signal Generator generates the clock and control signals for the ADCs and analog multiplexers on-chip as previously discussed in Section 3.3.3 and Section 3.4. The clock signals CK and CK_s serve as the ADC's main and auxiliary clocks respectively. The signals SMP, $STRT$, EN, control the operation of the ADCs, while the signals A_{1s}, A_{2s}, A_{3s}, A_{4s} control the operation of the analog multiplexers. These control signals are derived from an external 10-MHz crystal oscillator, which is mounted on the flexible PCB of the internal unit. The Data Packetizer takes as inputs the outputs from eight ADCs, appends parity bit to each sample, and packetize the data into a data frame before serially shifting the data off-chip to the on-board FPGA. The serial data is streamed out through ser_out pin of the Data Packetizer in Fig. 3-33. For further flexibility in the FPGA design, a parallel version of the data is also available through the 8-bit par_out pin of the Data Packetizer. For this purpose, the 5-bit channel address (chan_addr) and the clock for streaming parallel data off-chip (par_clk) are provided.

Control Signal Generator

Figure 3-34 shows a block diagram schematic of the Control Signal Generator. The main state machine of the Digital Control Unit is the 9-bit up-counter that keeps counting continuously from 0-319. The input clock of the counter is the 10-MHz external clock from the on-board crystal oscillator. One round of counting from 0-319 establishes the sampling period of each recording channel. Since one round of counting corresponds to 320 clock cycles, with the clock period of 100 ns, each recording channel is sampled at a sampling period of 32 μs. The 9-bit output of the counter (Q[8:0]) is fed to control logic blocks where it is used to generate the control

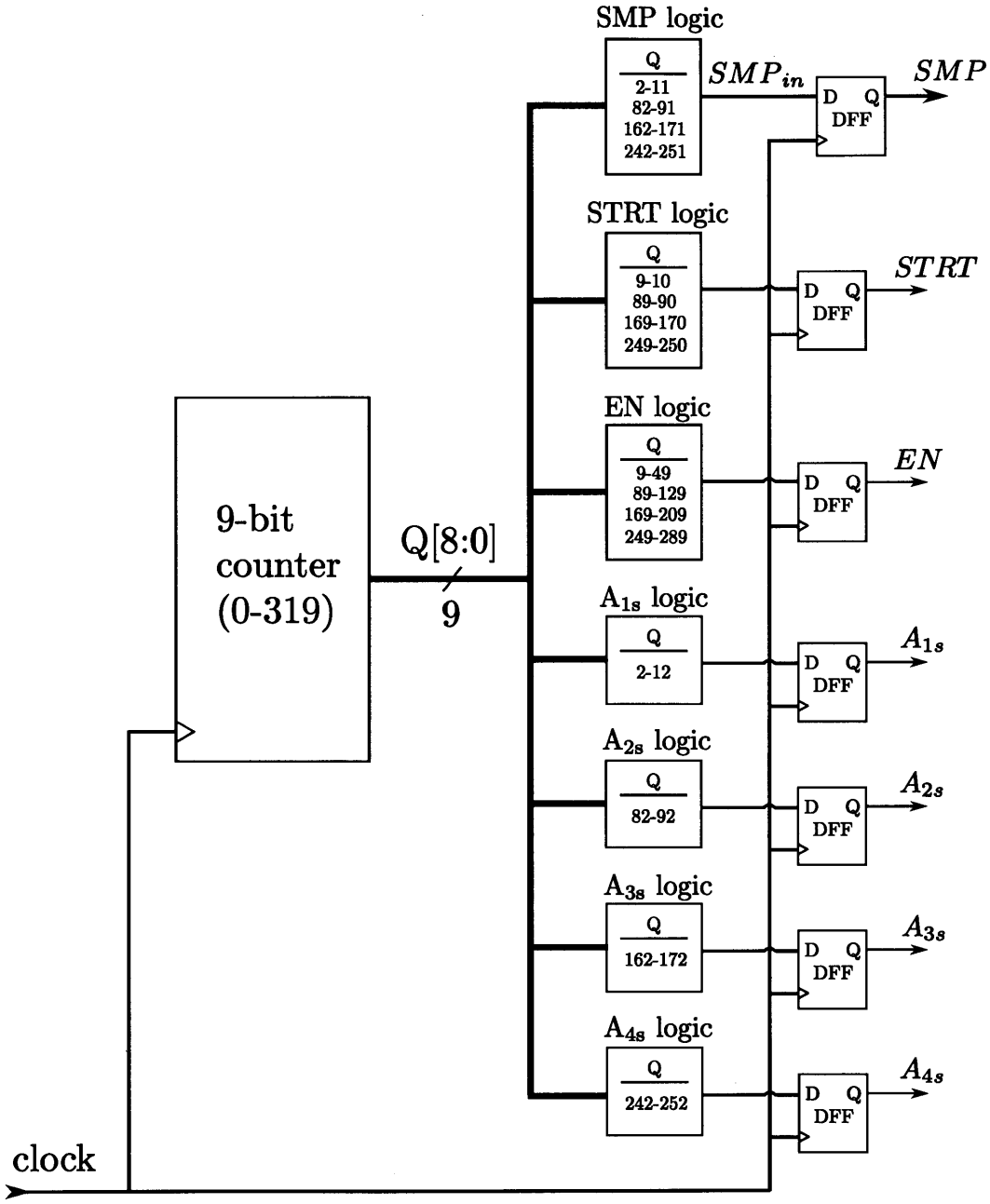


Figure 3-34: Block diagram of the Control Signal Generator.

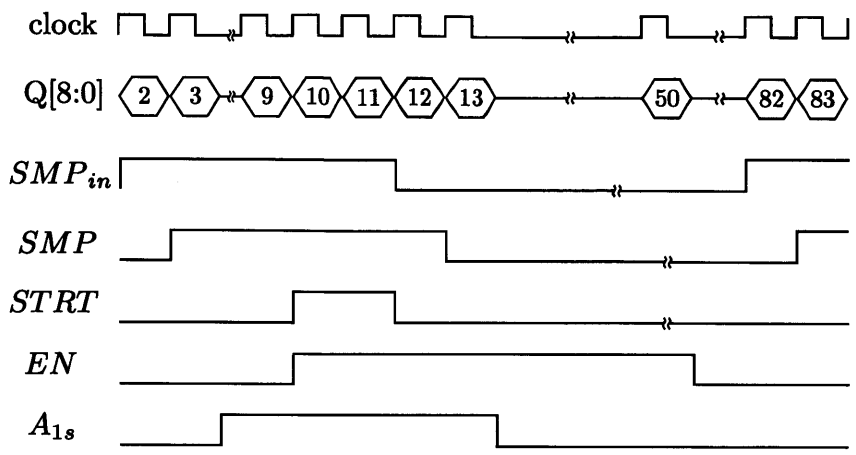


Figure 3-35: Timing diagram of Digital Control Unit. Q[8:0] is shown in decimal basis.

signals. For each control logic block in Fig. 3-34, the values listed under Q are the values of Q[8:0] in decimal basis at which the logic block asserts a 1. For example, the SMP logic asserts a 1 when the counter's output are 2-11, 82-91, 162-171, and 242-251. Note that there are four intervals in the whole 32 μs period at which SMP logic, STRT logic, and EN logic assert a 1 because each ADC must sample and convert the amplified neural signals four times from four neural amplifiers in a recording module. However, A_{is} logic, $i \in \{1, ..., 4\}$ only asserts a 1 once in the 32 μs period. The outputs of these 7 control logic blocks are then registered at the positive edge of the clock in order to create glitch-free control signals. Figure 3-35 shows the timing diagram of the control signals including the output of the counter (Q[8:0]) in decimal basis for the first quarter of the 32 μs sampling period. Note that SMP, $STRT$, EN, and A_{1s} are registered outputs, thus there is one clock period delay between when the outputs of the logic blocks go high and when these control signals go high. This scenario is illustrated in the SMP signal case in which the control signal SMP lags the output of its corresponding logic block (SMP_{in}) by 1 clock period.

The Control Signal Generator is also responsible for generating the clocks, CK and CK_s of the ADCs. Referring to Fig. 3-24, CK is a 2.5 MHz (T_{CK}=400 ns) clock that controls the operation of the SAR logic, while CK_s, also operating at 2.5 MHz, is used for sampling the outputs of the comparator to ensure the correct timing

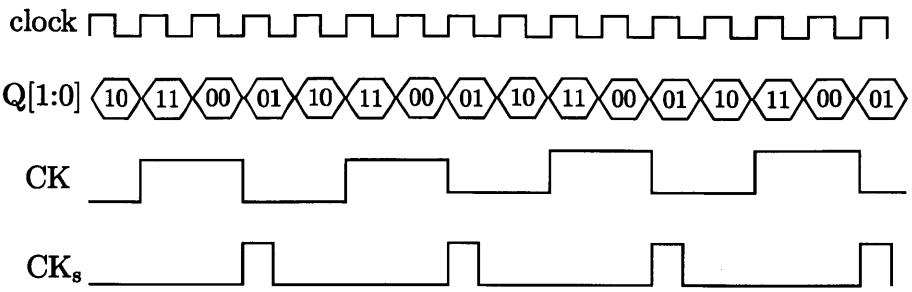


Figure 3-36: Timing diagram of the CK and CK_s generation. Q[1:0] is shown in binary basis.

operation of the ADCs. To generate CK and CK_s, we use the two least significant bits of the counter's output, Q[1:0]. The CK and CK_s are also registered to ensure that they are glitch-free. The timing diagram illustrating the generation of CK and CK_s is shown in Fig. 3-36. In Fig. 3-36, Q[1:0] is shown in binary basis.

Data Packetizer

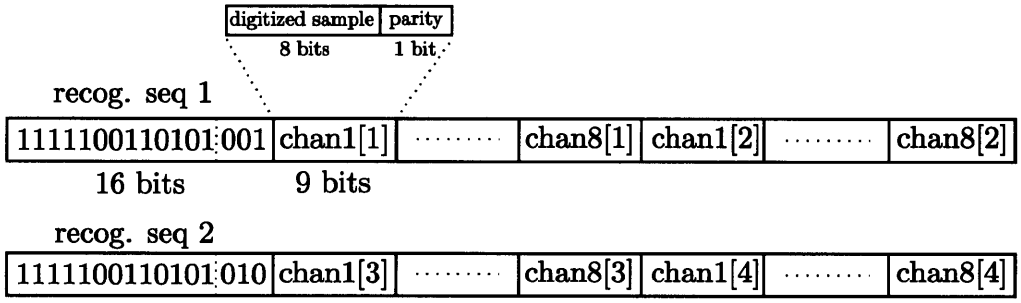


Figure 3-37: Format of the outgoing data packet of the 32-channel neural recording system.

With a 32 μs sampling period per channel and no storage mechanism on chip, the data from each recording channel must be transmitted off-chip within the next 32 μs. The Data Packetizer must collect all the outputs of the ADCs, organize these data into a data packet, and stream it serially off-chip, all within the 32 μs time frame. Serial streaming is also necessary if the 32-channel neural recording IC is to be interfaced with a wireless digital transmitter directly (for the subsequent generations of the recording system). To provide an easy synchronization between the 32-channel neural recording IC and the on-board FPGA, the data from all 32 recording channels

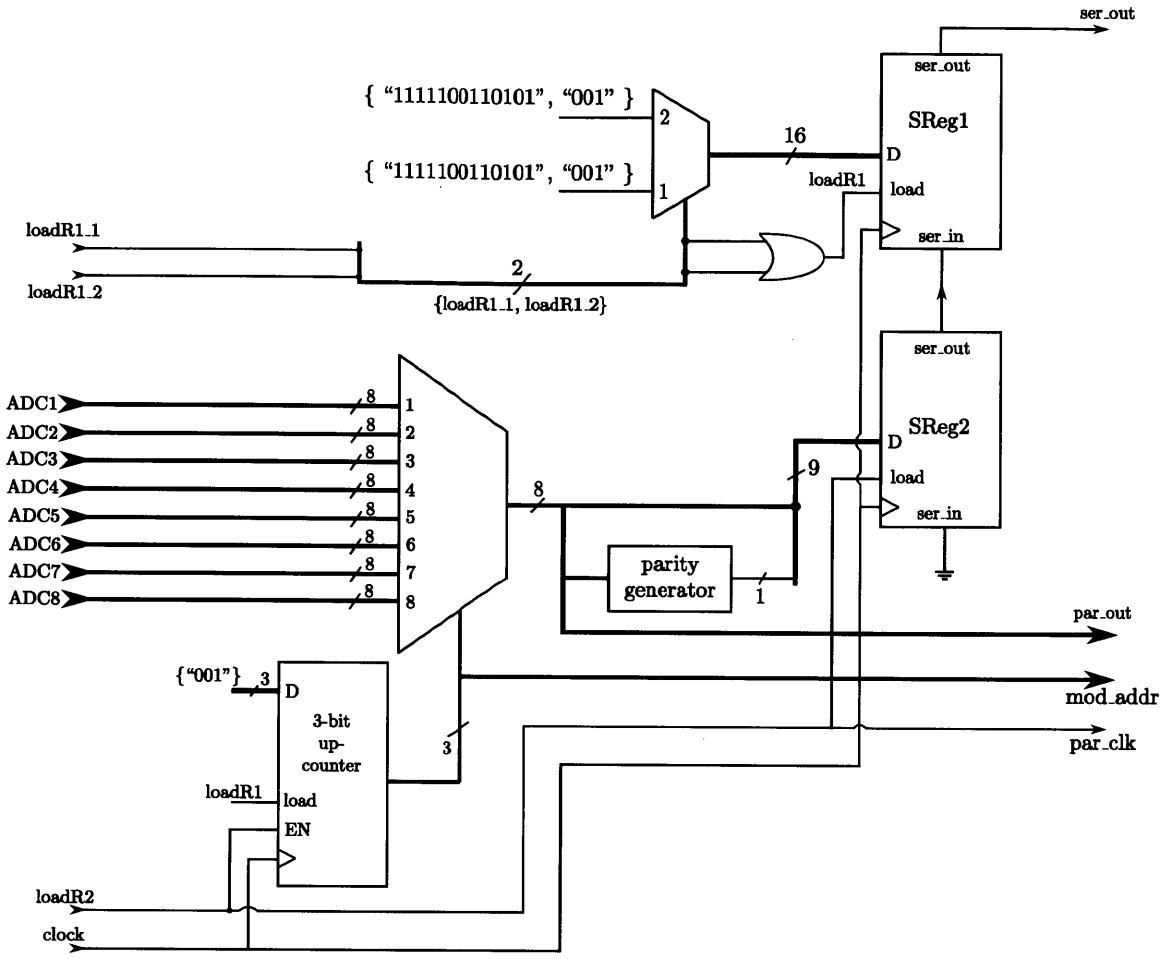


Figure 3-38: Block Diagram of the Data Packetizer.

are packetized into a 320-bit data frame. The data frame is streamed out serially at 10 Mbps and contains two 16-bit recognition sequences. The format of the outgoing data packet is shown in Fig 3-37. For error correction purposes, an even parity bit is appended to each digitized sample to create a 9-bit samples. In Fig. 3-37, the data frame is divided into two halves with each half containing one 16-bit recognition sequence and sixteen 9-bit samples from 16 recording channels. In Fig. 3-37, the 9-bit sample chan i[j] represents the sample from the recording module i ($i \in \{1, ...8\}$) and channel j ($j \in \{1, ..., 4\}$) respectively.

Figure 3-38 shows the schematic of the Data Packetizer that organizes the digitized data into data packets and streams the data serially off-chip. The main serialization process is performed by two parallel-in-serial-out shift registers SReg1 and SReg2.

Note that these two shift registers also contain serial inputs. The two shift registers are connected in a serial manner as shown in Fig 3-38. For each shift register, when the load pin is asserted, the shift register loads the parallel input upon the positive edge of clock. When the load pin is not asserted, the shift register shifts out its data content, also upon the positive edge of the clock.

The Data Packetizer requires three control inputs, loadR1_1, loadR1_2, and loadR2, which are generated from the 9-bit counter in the Control Signal Generator. These control inputs indicates when to load the content of each shift register such that the serial output stream is in the format as shown in Fig. 3-37. During operation, SReg1 and SReg2 will perform the data shifting function most of the time, except when these three control inputs are asserted. When the signal loadR1_1 is asserted, the 16-bit recognition sequence 1 ("1111100110101001") is loaded into SReg1. Similarly, when the signal loadR1_2 is asserted, the 16-bit recognition sequence 2 ("1111100110101010") is loaded into SReg1. The duration between loadR1_1 and loadR1_2 is 160 clock cycles apart, which corresponds to exactly the time required to stream out half of the 320-bit data frame. Loading parallel data into SReg2 is done more frequently than loading into SReg1, because when SReg2 is empty, the new sample needs to be loaded to ensure the continuity of streamed-out data. When load_R2 is asserted, a 9-bit sample from an ADC (8-bit digitized data + 1-bit parity) is loaded into SReg2. Therefore, loading the data into SReg2 happens every 9 clock cycles, except for the last sample right before the beginning of the next half of the data frame. After loading this particular sample into SReg2, we need to wait until SReg1 is empty before we can load the new recognition sequence into SReg1 and the new sample into SReg2. As a result, after the 9-bit samples for chan8[2] or chan8[4] are loaded (referred to Fig. 3-37), we need to wait for 9+16=25 clock cycles, instead of 9, before we can assert loadR2.

The timing diagram of the loading signals is given in Fig. 3-39. In this figure, the output Q[8:0] of the counter is shown in decimal basis. The signal loadR1_1 is asserted when Q[8:0]=71 while loadR1_2 is asserted when Q[8:0]=231, which is 160 clock cycles after loadR1_1 is asserted. The signal loadR2 is asserted when Q[8:0] is

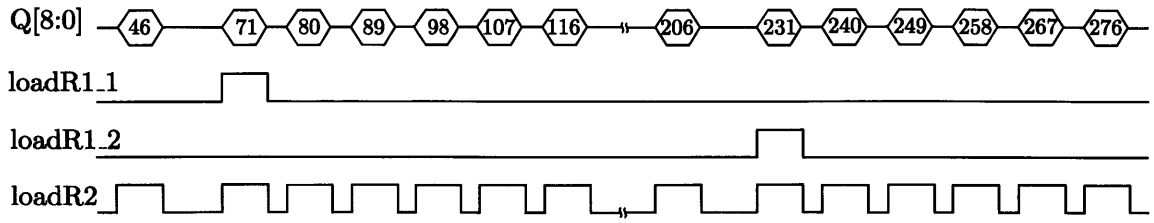


Figure 3-39: Timing diagram of the loading signals for loading SReg1 and SReg2.

equal to 71, 80, 89, ..., namely every 9 clock cycles. However, after loadR2 is asserted when Q[8:0]=206, it has to wait until Q=231 for it to be asserted again. As mentioned earlier, this is the 25 clock cycles required to empty the data content of SReg1, before the recognition sequence 2 can be loaded into SReg1. This same explanation also applies to when loadR2 is asserted at Q[8:0]=46. In this case, we need to wait for 25 clock cycles until Q[8:0]=71 before SReg1 is empty such that we can load recognition sequence 1 into SReg1.

3.7 Experimental Results

The 32-channel neural recording system was fabricated in the IBM 0.18 μm CMOS (7RF) technolog through MOSIS. The micrograph of the chip is shown in Fig. 3-40. Excluding the area for the I/O pads, the chip features the dimensions of 2.1 mm $\times$ 2mm. Due to the number of I/O pads included for testing purposes, the layout of the whole system was not optimized for the total chip area. However, the neural amplifier, analog multiplexer, and the ADC were laid out as compact as possible for further scaling to higher channel count in the subsequent generations. The active areas of the neural amplifier, the analog multiplexer, and the ADC are 0.03mm^2, 0.006mm^2, and 0.02mm^2 respectively. The remaining area of the recording module is occupied by the serial programming interface unit and the power supply decoupling capacitors.

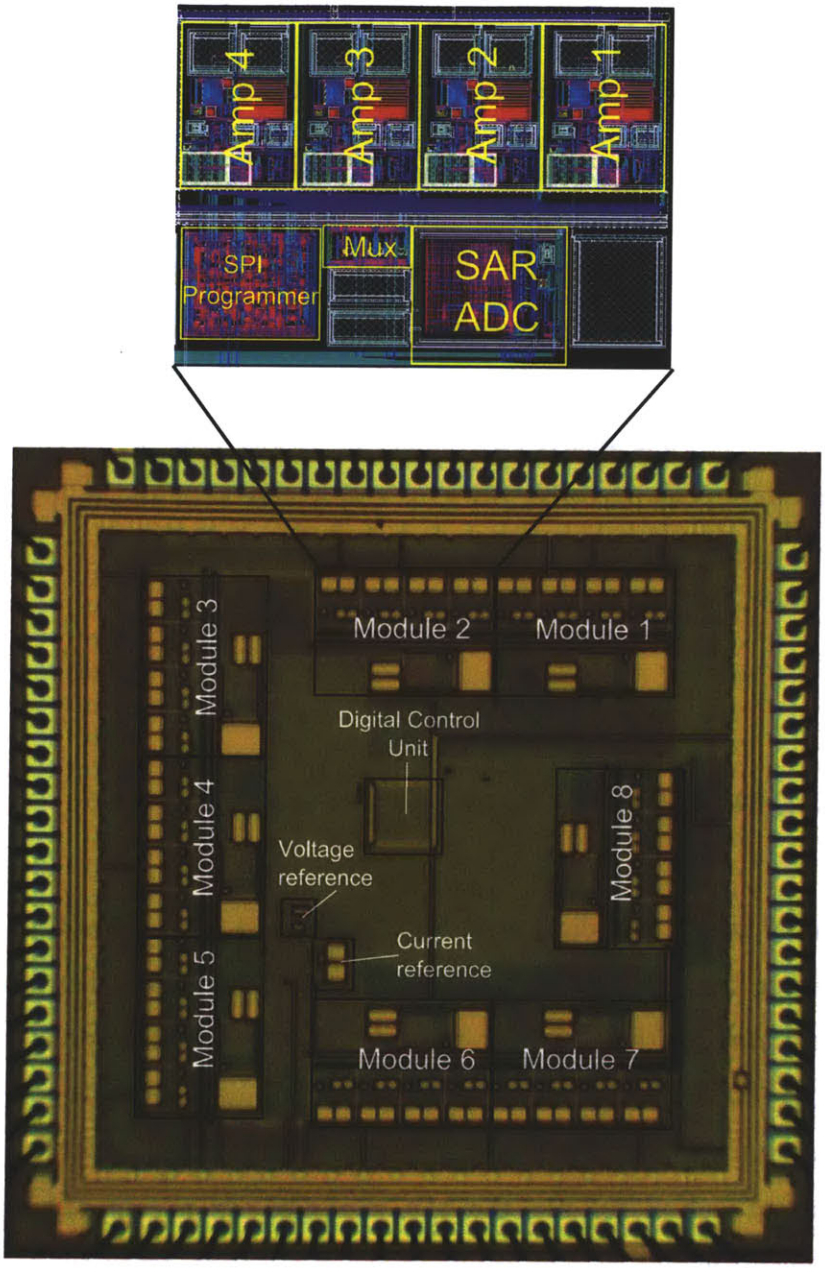


Figure 3-40: The micrograph of the 32-channel neural recording system.

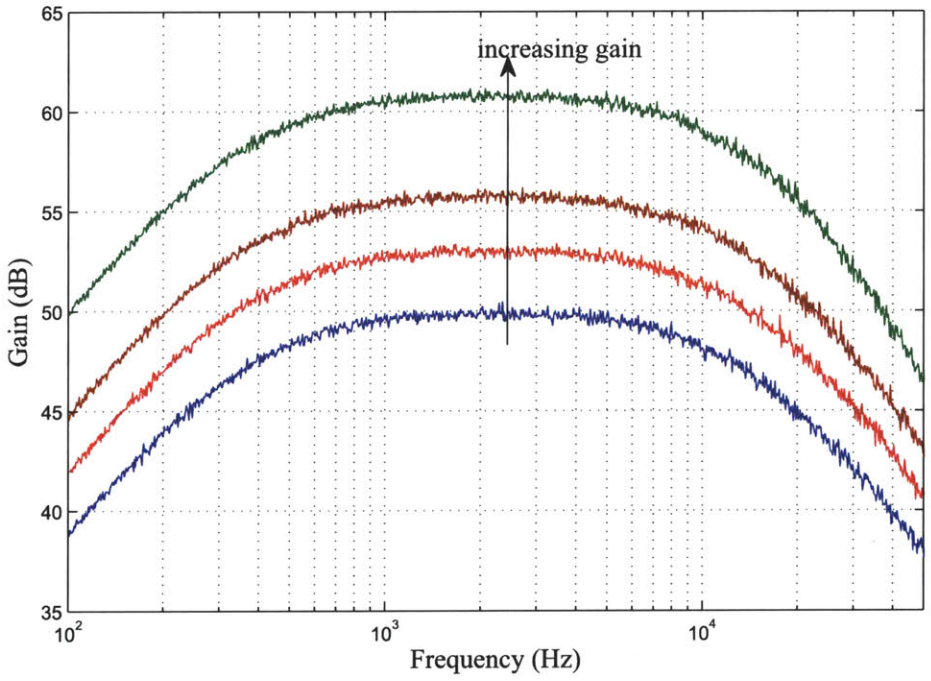


Figure 3-41: Magnitude Responses of the amplifier at different gain settings in the spike recording setting.

3.7.1 Benchtop Testing of the Neural Amplifier

In this section, we describe the experimental measurement of the neural amplifier. A dynamic signal analyzer (SR785, Stanford Research Systems) was used to measure the frequency responses and the output noise power spectral densities (PSD) of the neural amplifier. To measure frequency response, swept sine measurements were performed with an input amplitude of 100 μV. Figure 3-41 shows the measured magnitude responses of the neural amplifier in the spike recording setting for four different gain settings. The lower and upper -3 dB cutoff frequencies were measured to be $f_l = 350$ Hz and $f_h = 11.7$ kHz respectively, and were constant across different gain settings. The common-mode-rejection-ratio (CMRR) and the power-supply-rejection-ratio (PSRR) were measured to be 62 dB and 72 dB respectively. Figure 3-42 shows the input-referred noise spectral densities of the neural amplifier in the spike recording setting as we increased the front-end amplifier's bias current. The input-referred noise spectral densities were calculated by dividing the output noise spectral densities by the corresponding midband gains of the neural amplifier. Notice that as we increased the front-end amplifier's bias current, the input-referred noise spectral density in the frequency range above 100 Hz decreased as expected. However, at the frequency below 100 Hz, the input-referred noise spectral density was invariant to the front-end amplifier's bias current. At low frequency, the noise of the front-end amplifier was no longer the dominant factor because the gain from the front-end amplifier to the output of the neural amplifier dropped significantly due to the filtering effect of the bandpass filter ($f_l = 350$ Hz). As a result, the noise from the programmable gain amplifier (especially 1/f noise) became the dominant noise source of the overall neural amplifier and was independent of the front-end amplifier's bias current.

The total input-referred root-mean-square (rms) noise of the neural amplifier for different front-end amplifier's bias current levels (including the current from the local bias circuits) was calculated by integrating the corresponding input-referred noise density curve from 10 Hz to 65 kHz. Note that the bandwidth of integration (10 Hz-65 kHz) was several times the neural amplifier's bandwidth (350 Hz-11.7 kHz).

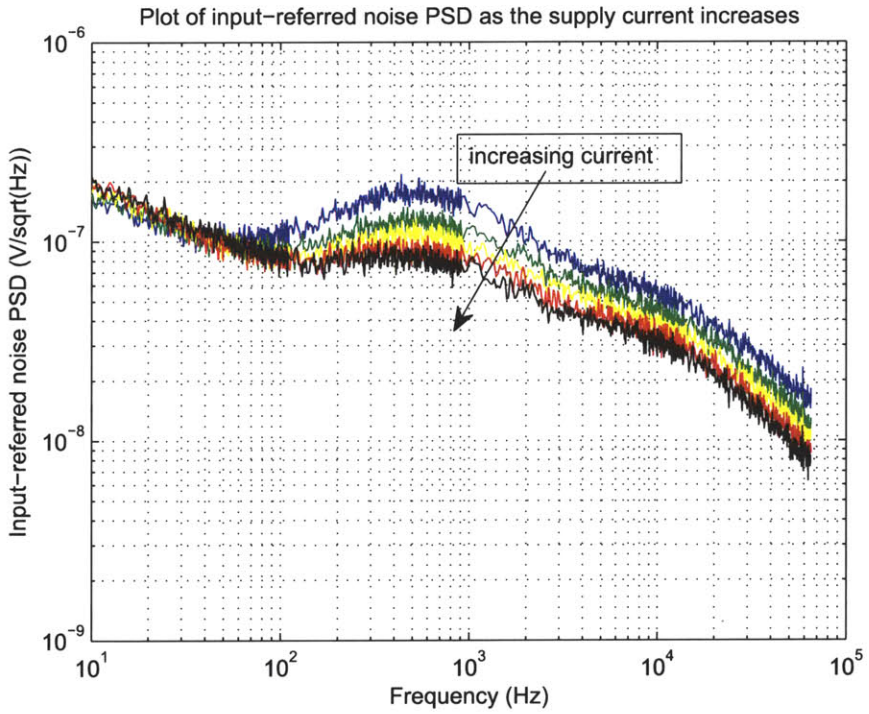


Figure 3-42: Input-referred noise densities of the neural amplifier as the front-end amplifier's supply current increases.

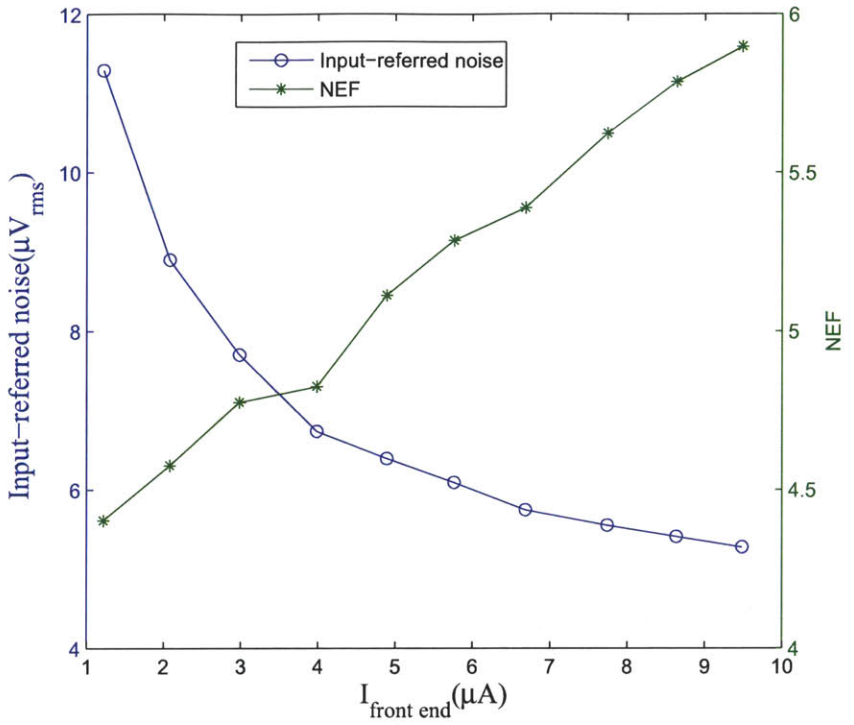


Figure 3-43: Integrated input-referred rms noise and the NEF of the neural amplifier vs. front-end amplifier's bias current.

The Noise Efficiency Factor (NEF) [63] was calculated for each front-end amplifier's bias current level. The plot showing the input-referred rms noise and the NEF versus the front-end amplifier's bias current level is shown in Fig. 3-43. It is interesting to note that as the front-end amplifier's bias current increased, its input-referred noise decreased as expected, however, at a slower rate. As we significantly increased the front-end amplifier's bias current, its thermal noise component decreased and the 1/f noise component became the limiting factor, thus limiting the improvement in the overall input-referred noise of the amplifier through the adaptive biasing technique. Our neural amplifier achieved the best NEF of 4.4 when the front-end amplifier's bias current was 1.2 μA. Due to the limitation from the 1/f noise discussed earlier, the NEF became worse as we increased the front-end amplifier's bias current.

As previously mentioned, the neural amplifier can also be configured to record LFPs. Figure 3-44 shows the magnitude response of the neural amplifier when con-

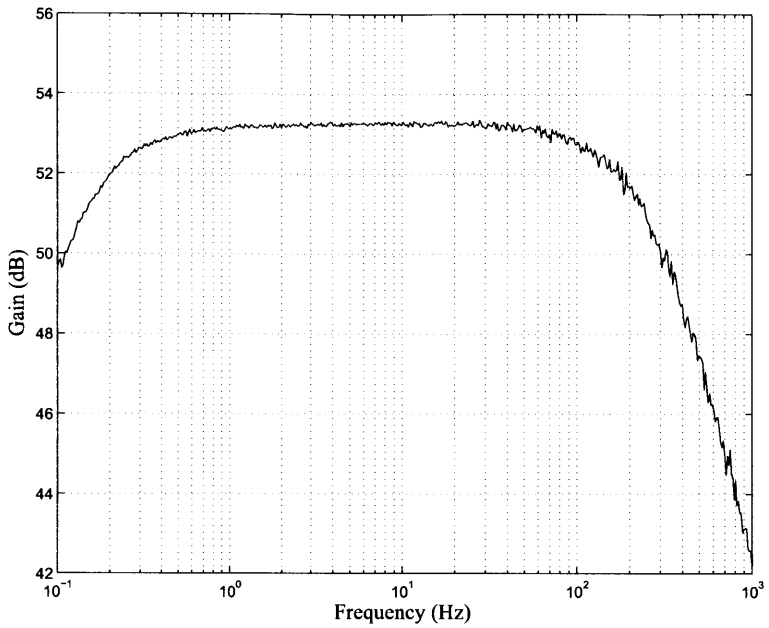


Figure 3-44: Magnitude Response of the amplifier in the LFP recording setting.

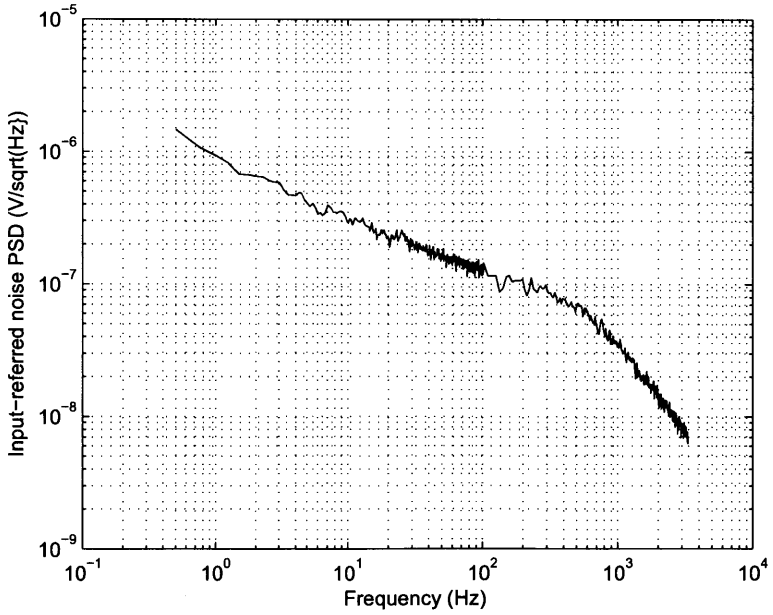


Figure 3-45: Input-referred noise density in the LFP recording setting

figured for the LFP recording setting. At this particular setting, the amplifier exhibits the midband gain of 53.3 dB with the -3 dB cutoff frequencies of f_l = 126 mHz and f_h = 293 Hz. The input-referred noise spectral density of the neural amplifier in the LFP recording setting is shown in Fig. 3-45. Integrating the input-referred noise spectral density curve from 500 mHz to 3.3 kHz yielded the total input-referred noise of 3.14 μV_{rms} with the front-end amplifier's bias current of 3.98 μA. At this low frequency, 1/f noise of the front-end amplifier dominates the overall input-referred noise of the neural amplifier making the adaptive biasing strategy impractical. Table 3.3 summarizes the measured performance of the neural amplifier.

Table 3.3: Performance Summary of the Neural Amplifier

Performance Metric	**Value**
Supply Voltage	1.8 V
Programmable Gain	49-66 dB
Bandwidth:	
spike recording setting	350 Hz-11.7 kHz
LFP recording setting	126 mHz-293 Hz
Input-referred noise (spike recording)	5.4 μV_{rms}-11.2 μV_{rms}
CMRR	62 dB
PSRR	72 dB
Power Consumption per Channel:	
Bandpass filter & Current DAC	700 nW
Programmable-gain amplifier	2.6 μW
Front-end amplifier	2.1 - 16.6 μW
Total	5.4 - 20 μW
Active Area	0.03 mm^2
NEF of 1^{st} stage	4.4-5.9

3.7.2 Benchtop testing of the Analog-to-Digital Converter

In this section, we describe the experimental characterization of the ADC. The static measurement was done with the code density test (histogram test) to obtain the integral nonlinearity (INL) and the differential nonlinearity (DNL) plots. To obtain a histogram, a rail-to-rail low-frequency sawthooth wave was fed into the input of the ADC. For an ideal ADC, if the frequency of the sawtooth wave and the ADC's conversion frequency are uncorrelated and enough output samples are collected, the

samples will tend to be equally distributed among all the possible output codes of the ADC, resulting in a flat histogram with every code bin containing the same amount of samples. For a real ADC, the ratio of the number of samples in a code bin to the total number of samples collected is proportional to the code width in the ADC's transfer curve. Thus, we can use the histogram to construct the ADC's transfer curve from which we can calculate the INL and DNL. In our measurement, we used a 10-MHz crystal oscillator as an external timing reference to the ADC while the sawtooth input was obtain from a function generator with independent timing reference. The ADC performed a conversion at the rate of 125 kS/s, while the frequency of the sawtooth wave is 300 Hz. Thus, both ADC's conversion rate and the frequency of the input signal are uncorrelated, and thus the code density method can be applied. A histogram from the code density test, which contain a total number of 327,680 samples, is shown in Fig. 3-46. From the histogram in Fig. 3-46, an ADC's transfer characteristic was constructed and the INL and DNL were calculated. The INL and DNL plots of the ADC are shown in Fig. 3-47. The least-squared approximation was used to calculate the INL. Both the INL and DNL of our ADC are within $\pm$ 0.4 LSB.

To measure the dynamic performance of the ADC, we used it to sample a full-scale (1 V_{pp}) low-distortion 1.024-kHz input sine wave from a dynamic signal analyzer (Stanford Research System, SR785). The sampling rate of the ADC was set at 125 kS/s. The Fast Fourier-Transform (FFT) analysis was applied to 31,982 sample points to obtain the plot of the power-spectral density presented in Fig 3-48. The signal-to-noise-and-distortion-ratio (SNDR) was calculated from this plot to be 47.81 dB. The effective number of bits (ENOB) was calculated from the SNDR to be 7.65 bits. The spurious-free dynamic range was 60 dB which was limited by the fifth harmonic. The total power consumption of the ADC from a 1-V supply voltage for this particular dynamic measurement was 1.93 μW which can be divided as follows: 592 nW from the comparator, 1.13 μW from the custom SAR logic, and 210 nW from the capacitor DAC array. Note that this power consumption is for converting the output signals from four amplifiers in a recording module. The average power

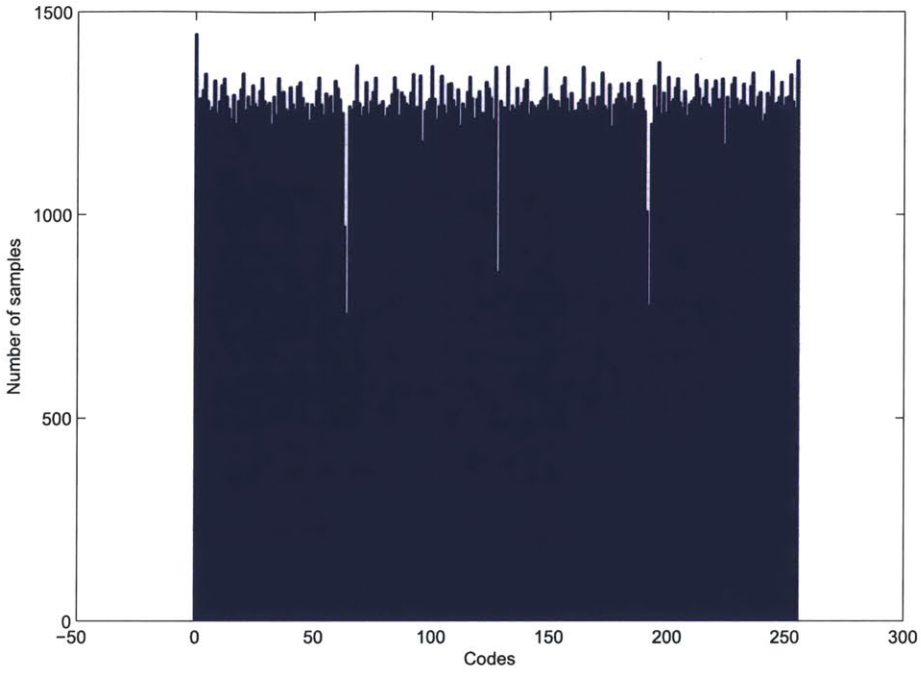


Figure 3-46: Histogram of the ADC's output codes from the code density test.

consumption per recording channel of the ADC is only 483 nW. One important figure of merit (FOM) of the ADC that is widely used to compare the ADCs across a wide range of bandwidths and precisions is the energy consumption per quantization level. The FOM can be calculated from the formula FOM $= P_{total}/\left(2^{\mathrm{ENOB}} \times f_s\right)$, where P_{total} and f_s are the total power consumption and the sampling frequency of the ADC respectively. The FOM of our ADC was calculated to be 77 fJ per quantization level and is among the most energy-efficient ADCs reported to date. The performance summary of the ADC is provided in Table 3.4.

3.7.3 Wireless *In-Vivo* Testing of the Neural Recording System in Behaving Primate

To verify the functionality of the neural recording system, we used it in our proof-of-concept prototype of the wireless recording system to obtain the neural data from a behaving primate. The recording unit was constructed on a custom PCB that in-

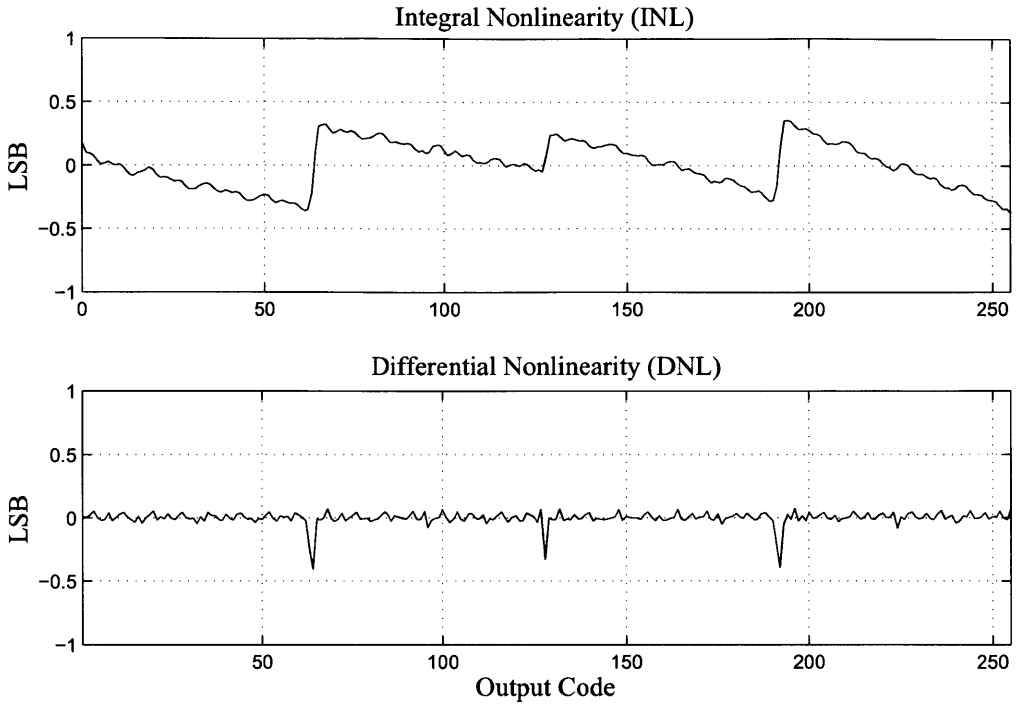


Figure 3-47: Low-frequency INL and DNL plots of our ADC. The INL is obtained using the least-squared approximation.

Table 3.4: Performance Summary of the Analog to Digital Converter

Performance Metric	**Value**
Supply Voltage	1 V
Full-scale voltage	1 V
Precision	8 bits
ADC's input range	0-1 V
Sampling rate	125 kHz
INL	$< \pm$ 0.4 LSB (8 bits)
DNL	$< \pm$ 0.4 LSB (8 bits)
SNDR	47.8 dB
SFDR	60 dB
ENOB	7.65 bits
Power Dissipation:	
Analog(comparator & Capacitor DAC)	802 nW
Digital (SAR logic)	1.13 μW
Energy per quantization level	77 fJ/State
Active Area	0.02 mm^2

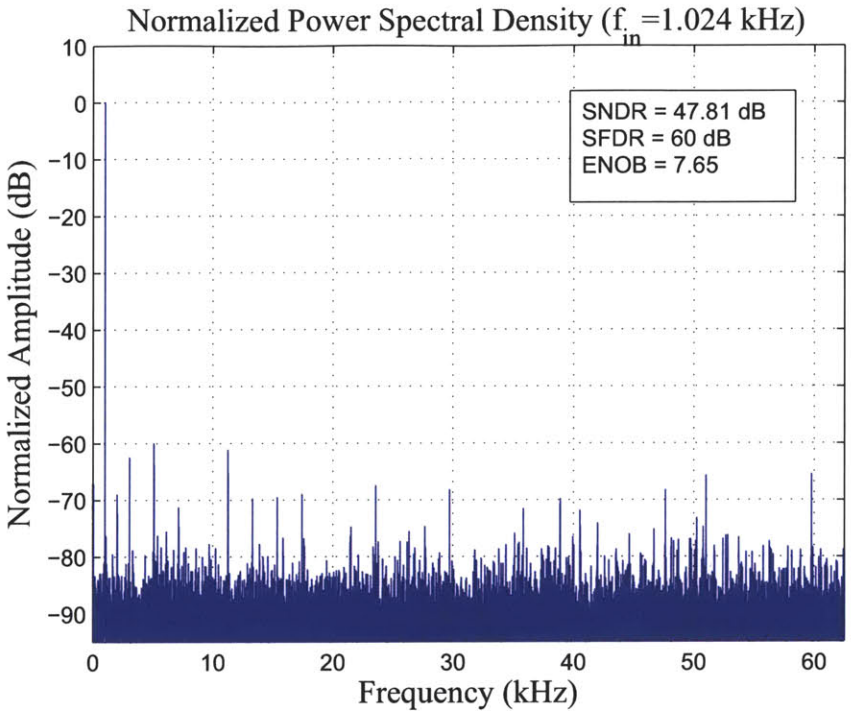
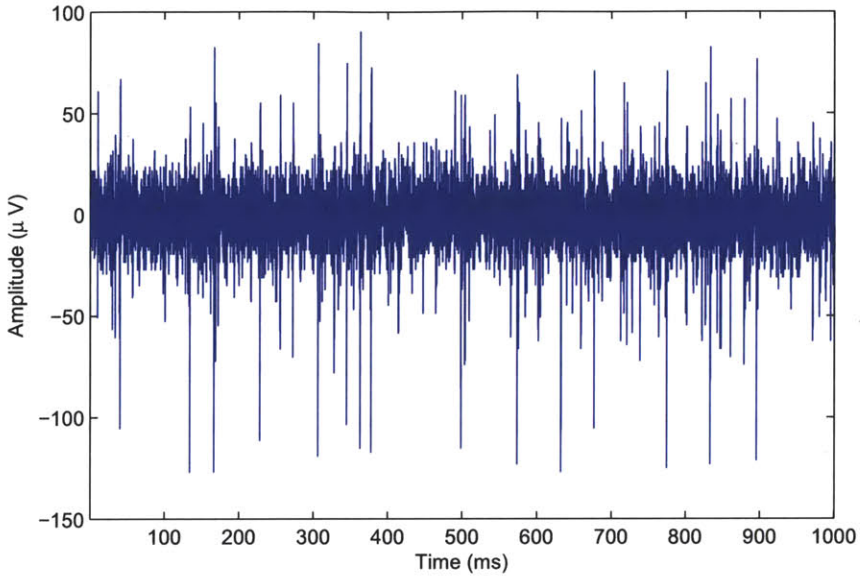
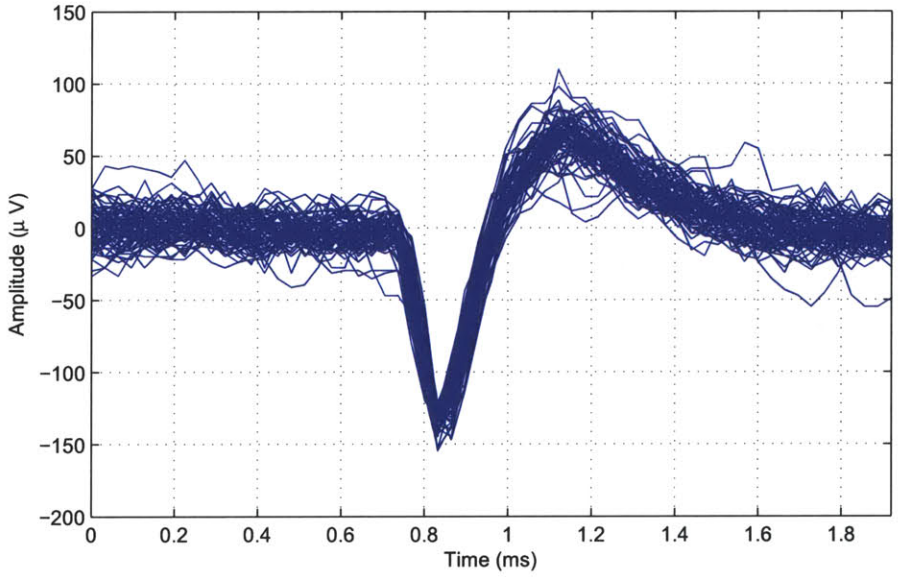


Figure 3-48: Measured output spectrum of the ADC with a rail-to-rail input sine wave of 1.024 kHz.

cluded the 32-channel neural recording chip (in QFP80 package), a low-power FPGA (IGLOO series, AGL060 from Microsemi), off-the-shelf power regulators, a 10-MHz crystal oscillator, the internal unit of the impedance-modulation data telemetry system described in [34], and the secondary coil. The receiver unit was constructed on another PCB that housed the external unit of the impedance-modulation data telemetry system and the primary coil. The receiver unit was interfaced with a PC via a USB data acquisition system (XEM3010 from Opal Kelly). The on-board FPGA in the recording unit selected the raw data from 8 input channels of the 32-channel neural recording system to be transmitted wirelessly to the receiver unit. Eight input channels of the recording unit were wired to microwire recording electrodes that had just been previously lowered into the brain tissue of a rhesus macaque while the rest of the input channels were grounded. Throughout the experiment, the primary and secondary coils of the data telemetry system were placed concentric to each other and spaced approximately 1 cm apart. Prior to each recording session, the configuration



(a) Long-time trace



(b) Superimposed spikes

Figure 3-49: Electrode-referred neural signals recorded from the brain of a rhesus macaque and transmitted wirelessly: (a) 1-second long raw neural data. (b) 74 superimposed spikes.

commands were sent wirelessly from the PC to the 32-channel neural recording chip on the recording unit. During the recording sessions, the recorded neural signals were transmitted at 2.5 Mbps from the recording unit to the receiver unit and subsequently streamed into the PC via a USB interface.

The wirelessly-recorded neural data from one of the recording channels is shown in Fig. 3-49. Fig 3-49(a) shows a 1-second long raw neural data that was streamed into the PC. Fig. 3-49(b) shows 74 neural spikes from this particular channel that have negative swings crossing the -120-μV threhsold, peak-aligned and superimposed on each other. The SNR of the recordings from our system was compared to the SNR of the recordings from a commercial system (Plexon, Inc) and both were found to be identical.

The performance summary of the neural recording chip during the wireless recording experiment is shown in Table 3.5.

Table 3.5: System Level Performance

Technology	0.18 μm CMOS
Voltage Supply:	
Amplifier array & reference circuits	1.8 V
ADC array & Digital Control Unit	1 V
Channel Count	32
Dimensions of the Neural Amplifier	215 μm $\times$ 155 μm
Dimensions of the Recording Module	680 μm $\times$ 450 μm
Die Dimensions	3.15 mm $\times$ 3.15mm
ADC's Input Range	0-1 V
ADC's Sampling Rate	31.25 kHz
INL of ADC	$< \pm$ 0.4 LSB
DNL of ADC	$< \pm$ 0.4 LSB
ENOB of ADC	7.65 bits
Power Dissipation:	
Neural Amplifier Array	207 μW (biased at NEF = 4.5)
ADC Array	15 μW
Digital Control Unit	42 μW
Voltage and Current References	61 μW
Total	325 μW

Table 3.6 compares the performance of the presented design with some of the state-

of-the-art designs in the literature that achieve low power consumptions and small areas per channel. The designs in [59] included both the recording and stimulation features while the design in [10] also included the digital signal processing (DSP) and ultra-wide-band (UWB) transmitter. In Table 3.6, only the recording features that include signal amplification and digitization are compared. Among recently reported designs with the comparable area per recording channel, our design achieved one of the lowest power consumption per channel reported to date.

Table 3.6: Comparison to other state-of-the-art neural recording systems

Reference	**[59]**	**[10]**	**[20]**	**This work**
Channel Count	128	128	16	32
Supply Voltage	3 V	±1.65 V	1.8 V	1.8 V(analog) 1 V (digital)
Technology	0.35-μm CMOS	0.35-μm CMOS	0.18-μm CMOS	0.18-μm CMOS
Mid-band gain	54-73 dB	57-60 dB	70 dB	49-66 dB
Low freq. cutoff	0.5-50 Hz	0.1-200 Hz	100 Hz	0.126 Hz, 350 Hz (selectable)
High freq. cutoff	500 Hz - 10 kHz	2 kHz - 20 kHz	9.2 kHz	293 Hz, 12 kHz (selectable)
Input-referred noise	6.08 μV_{rms} (10 Hz - 10 kHz)	4.9 μV_{rms} -	5.4 μV_{rms} -	5.4-11.2 μV_{rms} (10 Hz - 65 kHz)
NEF	5.5	-	4.9	4.4 - 5.9
ADC resolution	8 bits	6-9 bits (adjustable)	8 bits	8 bits
ADC sampling rate/channel	14 kHz	40 kHz	30 kHz	31.25 kHz
ADC INL & DNL	-	-	0.5/0.5	0.4/0.4
ADC ENOB	6.5 bits	-	7 bits	7.65 bits
Total Power Dissipation	2.43 mW	3 mW	680 μW	325 μW (@NEF=4.5)
Average Power /channel	19 μW/chan.	23.4 μW	42.5 μW/chan.	10.1 μW/chan. (@NEF=4.5)

3.8 Conclusion

In this chapter, we described the operation and the measured performance of the 32-channel neural recording IC, which is the heart of the implantable wireless neural recording system that will be the topic of the next chapter. The IC can amplify and convert the data from 32 input channels into 8-bit digital representations before sending the data off-chip in a serial-stream format. The adaptive biasing technique is utilized in the design of the neural amplifier to help minimize the total power consumption of the overall recording system. The measured performances of the circuit building blocks is presented. The neural amplifier is highly programmable; its gain, recording setting, and input-referred noise can be programmed to suit the recording environment, while occupying an area of only 0.03 mm^2. Depending on its input-referred noise level, the neural amplifier achieved an NEF in the range of 4.4-5.9 in the spike recording setting. The ADC achieved an ENOB of 7.65 bits while dissipating less than 500 nW per channel. The ADC's figure of merit of 77 fJ/State places it among the most energy-efficient ADC reported to date. The recording IC was successfully tested in an *in-vivo* wireless recording experiment from a behaving primate while dissipating only 10.1 μW/channel. Due to very small area and power consumption per recording channel, the recording IC is highly suitable for further scaling to higher channel counts in the subsequent generations.

Chapter 4

An Implantable Wireless Neural Recording System

In this chapter, I will describe the design of our implantable wireless neural recording system, which was introduced earlier in Chapter 1. Our wireless neural recording system consists of an internal unit and an external unit. The system is designed for use in a wireless recording experiment in a non-human primate. The internal unit is intended to be fully implanted under the skin (between the skull and the scalp) to eliminate any kind of transcutaneous connections. The external unit is intended to be placed over the top of the head to communicate with the internal unit and transfer power to operate the internal unit. While I will discuss the operation of the whole system, I will only focus on the design of the internal unit, especially on the signal processing aspect of the system.

4.1 System Overview

Figure 4-1 shows a block diagram of our wireless neural recording system. The internal unit contains 32 neural inputs which can be interfaced with 32 recording electrodes. The main function of the internal unit is to amplify and digitize neural signals from 32 recording electrodes, process the digitized neural data to obtain useful neural information, and transmit the neural information to the external unit via a short-

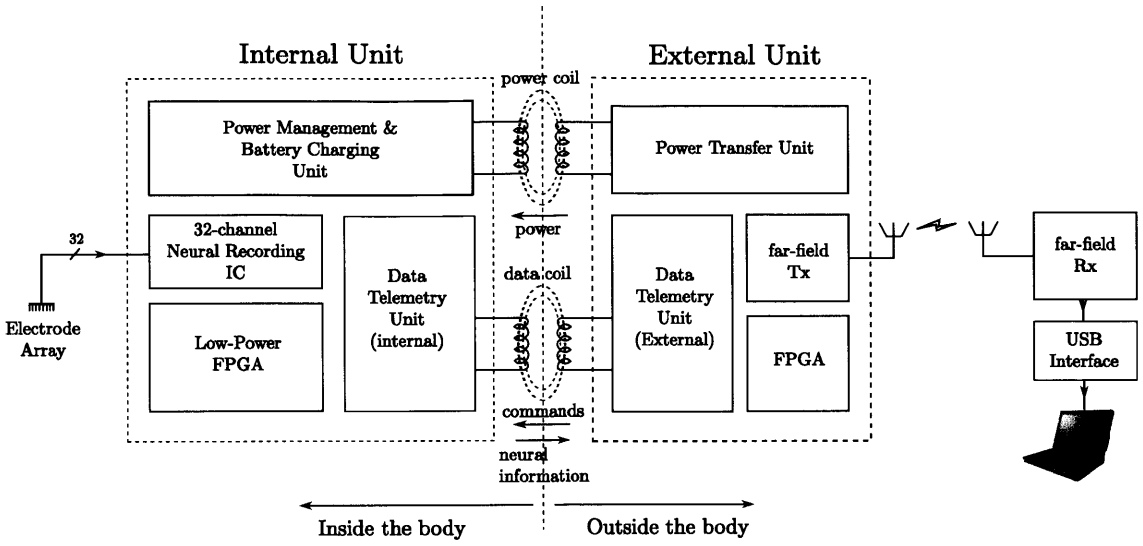


Figure 4-1: Block diagram of our implantable wireless neural recording system.

range wireless data telemetry system. The neural information from the internal unit is transferred inductively through the data coil in Fig. 4-1. The external unit receives the neural information from the internal unit, and then it can relay that information to a remote device via a far-field wireless transmission system. Since the power to operate the internal unit must be transferred wirelessly from the outside, the system also contains a power link through which RF power can be transferred from the external unit to the internal unit via the power coil as shown in Fig. 4-1.

4.2 Design Considerations of the Internal Unit

4.2.1 Energy Efficiency vs. Flexibility

Due to severe power constraint inside the body, our design goal is to minimize power dissipation of the internal unit. Besides the heat dissipation concern, if the internal unit is to be powered by an implantable battery, low power consumption of the internal unit could prolong the time between battery recharges, thus expanding the battery lifetime. An ideal strategy to minimize power consumption of the internal unit is to integrate all the functionalities into a single custom ASIC. With this approach, every circuit on the IC can be customized for a specific task such that the

overall system is highly energy-efficient. As mentioned in Chapter 1, such approach is only suitable for the final phase of system development when all the functionalities have been finalized. However, during the early phase of system development, some functionalities may still need to be modified or even completely redesigned, thus rendering the custom ASIC approach inappropriate. For design flexibility, we have adopted a multi-chip approach in which each part of the system is designed and optimized separately for performance and power consumption. The parts are then integrated at a PCB level to form a complete system. For added flexibility, an FPGA is utilized as the main signal processing and control units. The FPGA can be programmed *in-system* through the programming probe points included on the PCB. This *in-system* programmability gives us another level of flexibility in which the signal processing and control algorithms can be reconfigured during the testing phase of the hardware. Obviously, the drawback of this approach is a higher power consumption compared to that of the custom ASIC approach. Since the signals must be sent between ICs on the internal unit, each IC must incorporate output driver pads to drive relatively large parasitic capacitances associated with PCB wirings. Therefore, the chip-to-chip communication makes the power consumption of this multi-chip approach higher than that of the single chip approach. In addition, the FPGA is still much less energy-efficient than if we implement the same digital functionalities in a custom ASIC. However, at this phase of the system development, the multi-chip approach is chosen solely for design flexibility. Nevertheless, due to excellent energy-efficiency of our custom ASICs, the internal unit that will be presented in this chapter achieves one of the lowest power consumption among all the reported designs in the literature.

4.2.2 Power Minimization Strategies of the Internal Unit

In order to minimize power consumption of the internal unit, we utilize the following strategies. First, we use a highly energy-efficient 32-channel neural recording IC as a front-end processing stage. The design of the 32-channel neural recording IC is the subject of Chapter 3. In most designs, the front-end processing stage is the most

crucial part that determines the power consumption and performance of the neural recording system. Thus, it is very important that this front-end processing stage is very energy-efficient, while offering no compromise in performance. The second strategy for minimizing the power consumption of the internal unit is to keep the signal processing on the internal unit as simple as possible. Any computationally-intensive task should be performed outside the body (on the external unit) where power dissipation is less of a concern. This allows the signal processing on the internal unit to be performed with relatively low power consumption on a commercially-available low-power FPGA.

The third strategy for minimizing the power consumption in the internal unit is to minimize power dissipation of the wireless transmitter on the internal unit. Please note that the goal of our wireless neural recording system is to transmit neural information from the internal unit to a remote device. To minimize power dissipation of the wireless transmitter on the internal unit, we divide the wireless transmissions into two steps, instead of installing a powerful RF transmitter on the internal unit to send the neural information to the remote device all at once. The first transmission step is to transfer neural information from the internal unit to the external unit, which sits directly above the internal unit on the top of the head, via a short-range wireless data telemetry system. After receiving the neural information from the internal unit, the external unit can relay the neural information to a remote computer or a robotic limb via a separate far-field wireless transmission system. The short-range wireless data telemetry system that we use was presented in [34]. It consists of an internal data telemetry unit and an external data telemetry unit. The wireless data telemetry system utilizes an impedance-modulation technique to minimize power consumption of the internal data telemetry unit. The power consumed by the internal data telemetry unit can be as low as 100 μW, since this power is only used for turning on/off a switch that creates variation in a load on the internal data telemetry unit, which is then detected by circuitry on the external data telemetry unit. Since the external data telemetry unit consists of a free running oscillator and active circuitry to detect load variation on the internal data telemetry unit, its power consumption is

approximately 2.5 mW, which is much higher than that of the internal data telemetry unit. However, this power is dissipated outside the body where power sources are readily available. To transmit the neural information from the external unit to a remote device, a separate far-field transmission system can be used. Such far-field transmission system can be built from commercially available parts such as CC850 from Texas Instruments Inc. [5].

4.2.3 Powering the Internal Unit

Because the internal unit will be implanted under the skin to avoid any through-skin connections, the power to operate it must be transferred wirelessly from the external world. Inductive energy transfer has been widely studied for biomedical applications [71], [46], [30], [7]. In the neural recording applications, many systems have been presented that utilize continuous inductive power transfer to operate the internal units [25], [62], [49]. In this scheme, the RF power is continuously transferred to a receiving coil on the internal unit where the AC voltage across the receiving coil is rectified to create a DC voltage that is used for powering the internal unit. A supply filtering network must be included on the internal unit to smooth out the ripples on the supply rail. The sizes of the components of the supply filter are inversely proportional to the frequency of the RF carrier that is used for transferring power. To create a clean DC supply voltage with a small-sized supply filtering network, a high-frequency carrier must be used. However, due to elevated absorption by the body tissue at high frequency, high-frequency signal is not a viable solution for power transfer because as the frequency of the carrier increases, the body tissue absorption increases as well. This can lead to poor power transfer efficiency and overheating of the tissue that lies in the power transfer path. As a result, transcutaneous power transfer in biomedical applications normally utilizes low to moderate frequency (100 kHz-10 MHz range) to minimize losses in body tissue [55], [60], [18]. Thus, there exists a tradeoff between the size of the power supply network and the magnitude of voltage ripples on the power supply. In order to keep the size of the internal unit small, larger voltage ripples may need to be tolerated. Therefore, it is important

that the electronics on the internal unit exhibit good power supply rejection if the internal unit is to be continuously powered inductively. Otherwise, the sensitivity of the recording system can be severely degraded due to noise coupling from the power supply. An example of such problem was reported in [25]. In this work, the input-referred noise of an amplifier when it was powered in isolation from a clean power supply is 4.8 μV_{rms}. However, when the amplifier was used in the neural recording system that is powered inductively, the input-referred noise of that amplifier increased to 30-40 μV_{rms}.

One method to mitigate the noise problem from the power supply is to power the internal unit with an implantable battery. An implantable battery is also necessary if a wireless neural recording system is to operate without a top-of-the-head external unit (without continuous wireless energy transfer). Neural information may be transmitted directly from the internal unit to a remote host. An example of such system is reported in [11] where neural information is transmitted through an Ultra-Wide-Band (UWB) transmission system [37]. However, the work in [11] does not provide any detail of body tissue absorption of high-frequency signal (3.1-10.6 GHz) used by the UWB transmission system. Whether such UWB transmission is suitable for fully-implanted systems still requires further research.

Because a battery generates electrical energy from chemical reactions, the output voltage produced by a battery is much less noisy than the rectified voltage obtained from continuous power transfer. Even with a battery as a power source, inductive power transfer is still needed to periodically transfer energy to recharge the battery on the internal unit. Once the battery is successfully charged, the internal unit can operate continuously with a clean supply voltage. However, incorporating a battery on the internal unit does increase the size of the internal unit. How much increase in size depends on the size of the battery to be implanted, which is proportional to the amount of energy it can store. Therefore, internal units with high power consumptions are not normally powered by batteries [49], but are powered continuously by the inductive power transfer method instead. In our neural recording system, the power consumption of the internal unit is low enough such that battery powering is appli-

cable, and thus can be used as a viable method to improve neural recording quality. Therefore, our system is designed to be powered by an implantable battery and have incorporated a battery charging circuit [15] to periodically recharge the battery as needed.

4.2.4 Bandwidth Limitation of the Wireless Data Link

Since each recording channel of the 32-channel neural recording IC is sampled with 8-bit precision at 31.25 kHz, the overall data rate from the neural recording IC to the FPGA is (8 bits/chan. $\times$ 32 chan. $\times$ 31.25 kHz = 8 Mbps). If this amount of data is to be transmitted wirelessly to the external unit, a high-speed wireless transmitter needs to be used, thus significantly increasing the total power consumption of the overall internal unit. In addition, due to elevated body tissue absorption at high frequency, the transmitter becomes less efficient and needs to burn more power to compensate for signal attenuation by the body tissue. It is therefore advantageous to reduce the amount of data before wireless transmission such that a low-power wireless data link can be used. This would help keep the power consumption of the overall internal unit within a feasible limit.

Our wireless data telemetry system [34] can accommodate data transmission from the internal unit at a rate of 1 Mbps - 5 Mbps, with a higher data rate resulting in a higher bit-error rate (BER). To keep the BER low, we choose to transmit the neural information from the internal unit to the external unit at 2.5 Mbps, which is a rate at which our data telemetry system can comfortably accommodate. Therefore, data reduction needs to be performed such that the neural information can be transmitted wirelessly at 2.5 Mbps. A number of data reduction schemes have been proposed in wireless neural recording systems. The system in [26] utilizes an analog comparator, which is local to each amplifier, to perform spike detection on the neural amplifier's output. The spike threshold is set by a 6-bit current DAC. The comparator is reset roughly every 1 ms, resulting in a total data rate of 100 kbps for all 100 recording channels. To provide full visualization of a waveform, one channel out of possible 100 channels can be selected for full digitization at 10-bit precision. A more sophis-

ticated data reduction system was presented in [50]. This data reduction system was implemented in an FPGA and was designed for use in a 96-channel neural recording interface. The system provides four modes of the output data including: raw digitized data from one of the 96 channels, extracted waveforms of the spikes that cross threshold from any subset of the 96 channels, 1-ms bincounts of all the 96 channels, and raw digitized data from a single channel along with the extracted spike waveforms from that channel. To set the spike detection threshold on each channel, the system analyzes 16-ms length of the waveform from that channel and computes the appropriate threshold accordingly. The system in [20] implemented similar data reduction scheme as in [50] in an ASIC, and thus achieved much lower power consumption.

In our system, since we aim to minimize power consumption in the internal unit, a computationally-intensive scheme such as the one in [50] is avoided. Instead of focusing on transmitting raw neural data to the external unit, we aim to transmit decoded information from the internal unit to the external unit, which requires much lower transmission bandwidth. The decoded information from the internal unit can be used by a signal processing unit external to the body to derive motor control signals for controlling a prosthetic device. An energy-efficient neural decoding algorithm is implemented on the FPGA of the internal unit to derive the decoded information from the spiking information obtained from the 32-channel neural recording IC. The decoding algorithm only relies on simple additions, comparisons, and retrieving contents from small memories, and thus can be implemented with low power consumption on a small FPGA. The decoding algorithm and its implementation on an FPGA is beyond the scope of this thesis. For further detail, please see [48].

4.2.5 Interfacing between the 32-channel neural recording IC and the FPGA

As mentioned in Chapter 3, the 32-channel neural recording IC provides both the serial and parallel versions of the digitized neural data. For the serial version, the digitized neural data are packaged into a 320-bit data frame, which includes two 16-bit

recognition sequences to facilitate the synchronization between the IC and the FPGA. The data frame is streamed out serially at a rate of 10 Mbps. For the parallel version, the 32-channel neural recording IC provides 8-bit-wide bus that carries the digitized data, and a 5-bit-wide bus that carries the channel address. One output pin of the 32-channel neural recording IC provides a clock signal for streaming parallel data into the FPGA. The serial output version is necessary if the 32-channel neural recording IC is to be interfaced with a digital wireless transmitter directly. In addition, if many 32-channel neural recording ICs are to be integrated on a PCB to create a high-channel-count system, the serial output option can greatly reduce the number of traces that need to be routed on the PCB, thus significantly easing the PCB design process and reducing its fabrication cost. In this case, we need to implement logic circuits on the FPGA to demultiplex the digitized neural data from many serial pins into parallel buses before the neural data can be processed further by the FPGA. However, if only one 32-channel neural recording IC is to be integrated on a PCB, the parallel output version offers an advantage over its serial counter part. For the parallel output version, the FPGA can read the digitized neural data on the parallel bus directly without an additional demultiplexing logic block, thus saving the FPGA's resources for other important signal processing tasks. This can be critical for an FPGA with limited computational resources such as the one we are using in our system. Furthermore, incorporating a demultiplexing logic block, which must operate at the same speed as the incoming serial data (10 MHz), can increase the power consumption of the FPGA due to increased high-speed switching activity. On our PCB, both the parallel output bus and the serial output line are routed to the FPGA for design flexibility. Nevertheless, since our FPGA has limited computational resources and minimizing power consumption is our major design goal, we choose to use the parallel output version to avoid the need for an additional demultiplexing logic block on the FPGA.

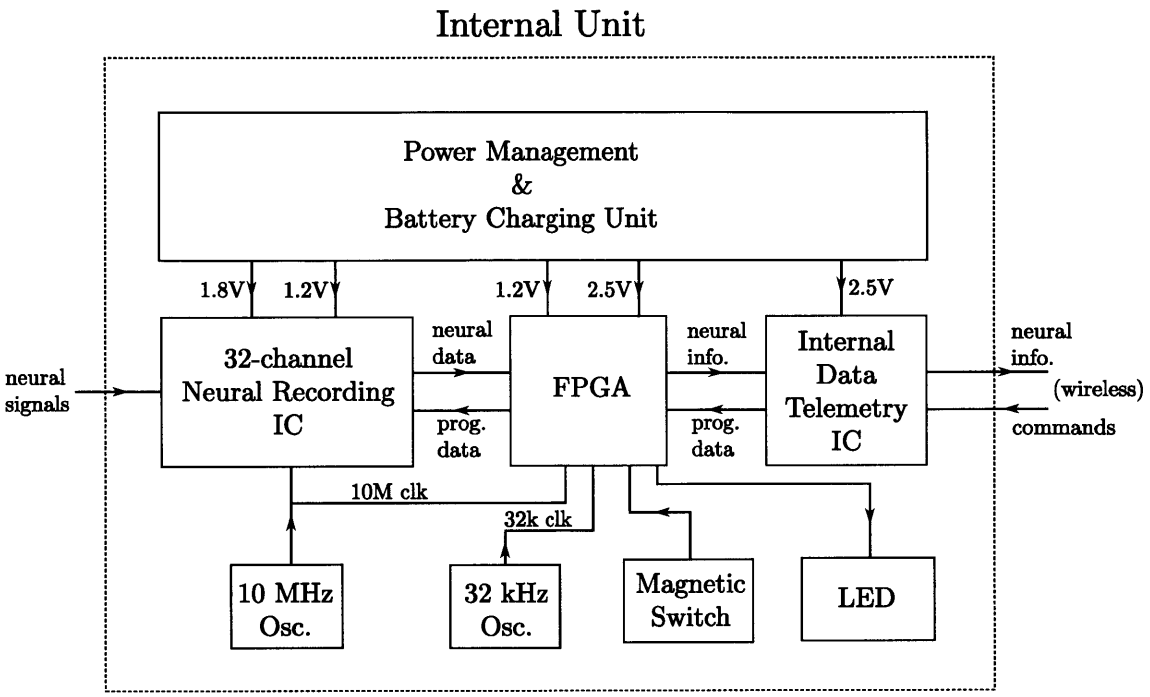


Figure 4-2: Block diagram of the Internal Unit.

4.3 Architecture of the Internal Unit

Figure 4-2 shows a detailed block diagram of the internal unit of our wireless neural recording system. The internal unit consists of: i) the 32-channel neural recording IC, ii) a low-power FPGA (IGLOO060V2, Microsemi Corp., Irvine, CA) [4], iii) internal data telemetry unit, and iv) power management & battery charging unit [15]. The internal unit also contains two crystal oscillators, a 10-MHz oscillator and a 32-kHz oscillator for clock generations, for frequency reference and a magnetic switch [2] to control the modes of transmission of the internal data telemetry unit, and a light emitting diode (LED) for battery charging status indication. The 32-channel neural recording IC is responsible for amplifying and digitizing neural signals from 32 recording electrodes. The digitized neural data is sent from the neural recording IC to an on-board low-power FPGA for further processing. The FPGA is packaged in a 121-pin chip-scaled package (CS121) with dimensions of 6 mm $\times$ 6 mm. The FPGA performs two important tasks including i) signal processing and data reduction, and ii) operation control of the internal unit. The processed neural information from the

FPGA is sent to the internal data telemetry unit to be transmitted wirelessly to the external unit (uplink telemetry mode). In addition, the internal data telemetry unit is responsible for receiving configuration commands from the external data telemetry unit (downlink telemetry mode). The commands are used for configuring various parameters of the internal unit. To switch between the uplink and downlink telemetry modes, a magnetic switch is included on the internal unit to configure the internal data telemetry unit into one of the transmission modes. Since the internal unit will be fully implanted under the skin, the power to operate it must be transferred wirelessly. Thus, the internal unit contains the power management & battery charging unit for receiving RF power from the external unit to recharge the implantable battery. The power management & battery charging unit is also responsible for generating different supply voltages to power various subsystems on the internal unit.

4.3.1 Power Supply Domains of the Internal Unit

The internal unit integrates many ICs from different technologies, thus they were designed to operate from different supply voltages. The 32-channel neural recording IC was designed to work with supply voltages of 1.8 V (analog) and 1 V (digital). The processing core of the FPGA can operate with a supply voltage from 1.2 V to 1.5 V, while many supply voltages can be used for its I/O. However, during the flash programming of the FPGA, a 1.5 V supply voltage is needed [4]. The internal data telemetry unit was designed to operate from a 2.5 V supply voltage. The 10 MHz crystal oscillator was designed to operate from a 1.8 V supply voltage, while the 32 kHz crystal oscillator was designed to operate from a supply voltage between 1.3-5.5 V. Due to these various requirements of the supply voltages, the internal unit needs to provide various power supply domains to operate each subsystem at its ideal condition.

For our internal unit, the Power Management & Battery Charging Unit generates three supply voltage domains including: i) 1.8 V supply domain, ii) 1.2 V supply domain, and iii) 2.5 V supply domain. The 1.8 V supply domain is used to power analog circuitry (neural amplifiers and multiplexers) in the 32-channel neural record-

ing IC and the on-board 10 MHz crystal oscillator. The 1.2 V supply domain is used to power the ADCs, the Digital Control Unit, the digital I/O pads of the 32-channel neural recording IC, the core of the FPGA, and the FPGA's I/O pads that interface with the 32-channel neural recording IC. The 2.5 V supply domain is used to power the internal data telemetry unit, the LED, the magnetic switch, the 32 kHz crystal oscillator, and some FPGA's I/O pads. Since during the flash programming of the FPGA, the core of FPGA must operate from a 1.5 V supply voltage, we therefore include a voltage switching circuit to switch the output voltage of the 1.2 V supply domain to 1.5 V. However, during normal operation of the internal unit, the 1.2 V supply voltage always generate the output of 1.2 V.

4.3.2 Design of the Digital System on the FPGA

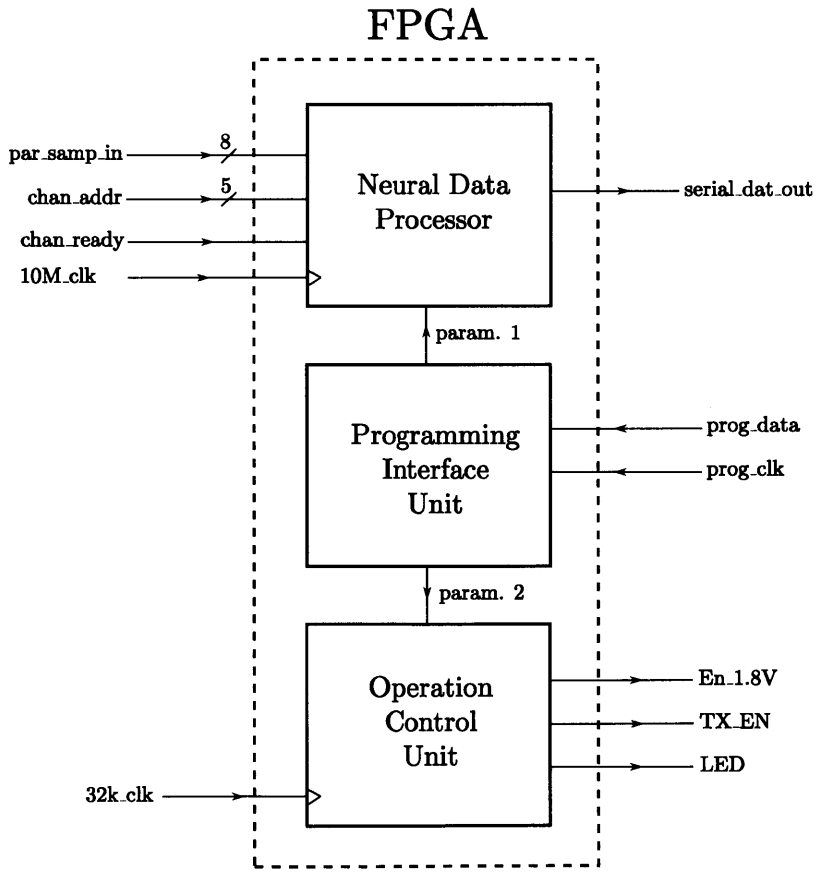


Figure 4-3: Block diagram of the synthesized system on the FPGA.

Figure 4-3 shows the overall architecture of our digital system on the FPGA. It consists of three main modules including the Neural Data Processor, the Operation Control Unit, and the Programming Interface Unit. The Neural Data Processor takes as inputs the digitized neural data (8-bit bus par_samp_in) and the channel address information (5-bit bus chan_addr) from the 32-channel neural recording IC. The signal chan_ready is a signal notifying whether the data on par_samp_in and chan_addr buses are ready to be used by the Neural Data Processor. The Neural Data Processor operates from the 10 MHz clock (10M_clk in Fig. 4-3), which is generated from the on-board 10 MHz crystal oscillator. It then processes the neural data on par_samp_in bus to obtain important neural information and reduce the amount of data for transmission. The processed neural information is then streamed out serially at a rate of 2.5 Mbps to the internal wireless data telemetry unit, where it is transmitted wirelessly to the external unit through the data coils. The operation of the Neural Data Processor will be explained in Section 4.4.

To process the neural data, the Neural Data Processor needs a certain parameters that must be provided by the user. As an example, the threshold of each recording channel must be set in order for the Neural Data Processor to detect the threshold crossing events correctly and accurately. Therefore, it is important that the user be able to wirelessly set the values of parameters on the FPGA. For this purpose, the Programming Interface Unit in Fig. 4-3 is included on the FPGA. The Programming Interface Unit is of similar architecture to the one explained in Section 3.5 of Chapter 3. The design of the Programming Interface Unit and the description of important parameters for the operation of the internal unit will be explained in Section 4.6.

For practical BMIs, the internal unit of the recording system is not just transmitting the neural information all the time. However, the user occasionally needs to instruct the internal unit into various modes of operation. For example, the user may want to turn off the internal unit to save battery power, or the user may want to reconfigure the parameters of the 32-channel neural recording IC or the spike detection thresholds of some recording channels. In addition, the user may want to change the way in which the neural information is being sent out (explained later in

Section 4.4). Or after a long recording session, the battery might be running out of charge and thus needs to be recharged. In this case, the user may wish to turn off most of the circuits on the internal unit such that all the charging current can be directed toward charging the implantable battery. Therefore, the internal unit needs a centralized control unit to control its operation based on the user's instructions. For this purpose, we include the Operation Control Unit on the FPGA to control the operation of the internal unit. The operation of the Operation Control Unit will be explained in Section 4.5.

4.3.3 Output Modes of the Neural Data Processor

The Neural Data Processor receives the digitized neural data from the 32-channel neural recording IC and then processes these neural data to reduce the amount of information that needs to be transmitted wirelessly to the external unit. The Neural Data Processor can output the neural information in one of the three modes including *Decode*, *Calibration*, and *Stream8* modes, which can be set by the user.

In BMI applications, the main signals of interest are not just the high-bandwidth raw neural data, but also the low-bandwidth decoded neural information that may be used to derive the motor control signals for a prosthetic device. Thus, it is beneficial if the internal unit can derive the decoded neural information and only transmit this information to the external unit. This method would greatly reduce the amount of data that needs to be transmitted wirelessly, thus lowering the power of the wireless data telemetry system, however, at an expense of slightly increased power consumption on the internal unit. In our system, we incorporate an energy-efficient neural decoding unit to perform such data reduction scheme. The detail of the neural decoding unit can be found in [48]. The neural decoding unit takes as inputs the threshold crossing information of spikes from all 32 channels. To provide the threshold crossing information to the neural decoding unit, a digital spike detection logic is implemented on the FPGA. Note that decoding operation can be done on the external unit as well. This is actually a more widely adopted approach since sophisticated algorithms can be implemented without as stringent a power constraint as that of the implantable

decoding unit. Performing neural decoding on the external unit also offers greater flexibility since the algorithms can be modified until a satisfactory performance is achieved, even after the internal unit has already been implanted. To provide such flexibility, we also provide the threshold crossing information, which is the input of the internal neural decoding unit, to the external unit. We term this first output mode of the Neural Data Processor as the *Decode* mode. In this mode, the output of the Neural Data Processor contains both the decoded neural information, and the threshold crossing information of neural spikes from all the 32 channels for external decoding tasks.

In order to accurately set the threshold of each recording channel for the spike detection logic, raw neural data from each channel is required for the computation of the baseline level and noise of each recording channel. To avoid performing computationally-intensive task on the internal FPGA during the threshold calculation process, we choose to calculate the spike detection thresholds and the baseline levels of all the 32 channels on an external processor. Therefore, the Neural Data Processor must be able to provide the raw waveform information from all 32 recording channels to the external unit for visualization by the user. Due to transmission bandwidth limitation of the wireless data link discussed in Section 4.2.4, the neural data from all the 32 channels cannot be transmitted to the external unit all at once. The wireless data link can only accommodate the raw neural data up to eight channels. However, to ease the threshold calculation process, the waveform information from all the 32 recording channels should be obtained in one recording session. For this purpose, we include another output mode for the Neural Data Processor, which we term the *Calibration* mode. In this mode, we divide the 32 recording channels into four nonoverlapping sets, with each set containing 8 recording channels. The *Calibration* session is then divided into 2-second time frames. The Neural Data Processor then rotates to each set of eight recording channels and transmits 2-second snippets of raw neural data from eight recording channels in that set. After at least 8 seconds of the *Calibration* session, the Neural Data Processor will manage to transmit at least 2-second length of raw neural data snippets from all 32 recording channels.

This amount of data on each channel should suffice for the computation of the noise amplitude and the baseline level of each recording channel. If a *Calibration* session of longer than 8 seconds is performed, the neural data snippets of each recording channel can be concatenated to provide more data for calibration purpose.

In experiments in which raw neural data is of primary concern such as in physical neuroscience studies, the user may want to obtain raw neural data from as many channels as possible. Therefore, we have included an output mode of the Neural Data Processor for streaming out raw neural data from many recording channels at once. As mentioned in Section 4.2.4, we choose to transmit the neural information from the internal unit to the external unit at a rate of 2.5 Mbps. This data rate is equivalent to sending ten 8-bit bytes in a 32 μs time frame. If only raw neural data are to be sent out in a 32 μs time frame, the wireless data link can accommodate up to 10 channels. However, to provide synchronization between the internal unit and the external unit, the data stream needs to contain extra bits for data synchronization purpose. Therefore, instead of transmitting only raw neural data from 10 channels, we choose to transmit only 8 channels of raw data in each 32 μs time frame, while the other two bytes are preserved for synchronization purpose and providing information on the status of the data being transmitted. The eight channels to be transmitted can be arbitrarily selected by the user through downlink wireless communication. We term this output mode *Stream8* mode. To choose between the three output modes, the user must wirelessly program the content of a 2-bit register "Rec_Mode." Table 4.1 summarizes the relationship between the register Rec_Mode and the output format of the Neural Data Processor.

Table 4.1: Rec_Mode vs. Output Format of the Neural Data Processor.

Rec_Mode	Mode	Description
01	*Calibration*	snippets of data from all 32 channels
10	*Stream8*	raw neural data from selected 8 channels
11	*Decode*	decoded neural information + threshold crossing information

4.4 Neural Data Processor

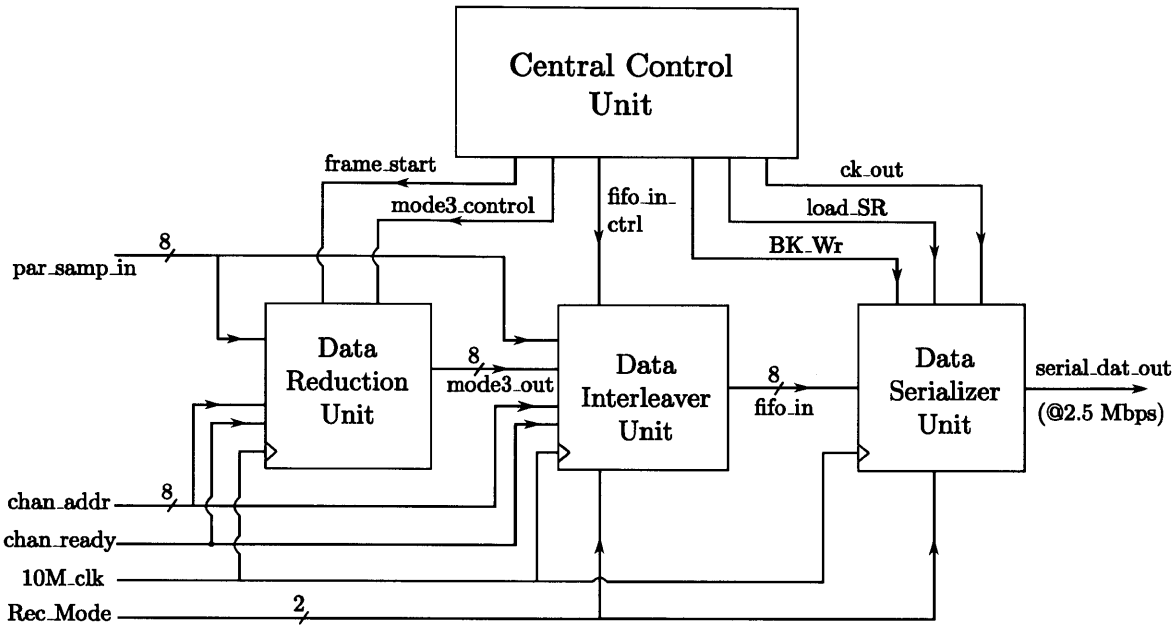


Figure 4-4: Block diagram of the Neural Data Processor implemented on the FPGA.

The block diagram schematic of the Neural Data Processor is shown in Fig. 4-4. The main processing blocks of the Neural Data Processor are the Data Reduction Unit, the Data Interleaver Unit, the Data Serializer Unit, and the Central Control Unit. The Data Reduction Unit performs the data reduction on the digitized neural data (on par_samp_in) by i) detecting the threshold crossings to obtain spiking information on each recording channel, and ii) performing the decode on the spiking information to obtain decoded neural information. The output of the Data Reduction Unit is denoted as mode3_output in Fig. 4-4, and consists of both the decoded neural information and the threshold crossing information, which are organized in a certain format. The operation of the Data Reduction Unit will be explained in Section 4.4.1. The output of the Data Reduction Unit is then sent to the Data Interleaver Unit. The Data Interleaver Unit is responsible for selecting what data to be passed to the Data Serializer Unit such that the serialized output data, serial_dat_out, in Fig. 4-4 is in a format recognized by the external unit. To specify the output format of the Neural Data Processor, the content of the register Rec_Mode must be set. If Rec_Mode =

"11" (in binary basis), the Data Interleaver Unit will select the output of the Data Reduction Unit (mode3_output) to be passed to the Data Serializer Unit. However, if Rec_Mode is either "01" or "10" (in binary basis), the Data Interleaver Unit will select digitized neural data from some recording channels to be passed to the Data Serializer Unit. The operation of the Data Interleaver Unit will be explained in Section 4.4.2. The Data Serializer Unit takes the parallel output bytes from the Data Interleaver Unit, serialize them, and send the serial data out to the internal wireless data telemetry system at a rate of 2.5 Mbps. The operation of the Data Serializer Unit will be explained in Section 4.4.3. The Central Control Unit is responsible for generating important timing control signals for the three processing units previously mentioned. The operation of the Central Control Unit of the Neural Data Processor will be discussed in Section 4.4.4.

4.4.1 Data Reduction Unit

Figure 4-5(a) shows the diagram of the Data Reduction Unit. It consists of the Neural Decoding Unit and the Spike Detection Logic. The operation of the Neural Decoding Unit is beyond the scope of this thesis, thus we will treat it as a signal processing block that takes as input the 32-bit spiking information from the Spike Detection Logic and gives as output the 32-bit decoded neural information. The main processing block of the Spike Detection Logic is the comparator (ThComp) that detects if the 8-bit neural data (par_samp_in) is greater than its corresponding 8-bit spike detection threshold (Th). To aid the description that follows, let's consider the timing diagram shown in Fig. 4-5(b). The signal chan_addr is a 5-bit signal that provides the channel address information of the neural data on par_samp_in. In Fig. 4-5(b), chan_addr is presented in decimal basis. The recording channels are numbered as channel 0 to channel 31. The signal chan_ready is a timing reference signal to notify when the data on par_samp_in and chan_addr are ready to be retrieved. The signal frame_start, which is generated from the Central Control Unit of the Neural Data Processor, serves as the timing reference that specifies the beginning of the 32 μs data frame which contains the neural data from all 32 channels.

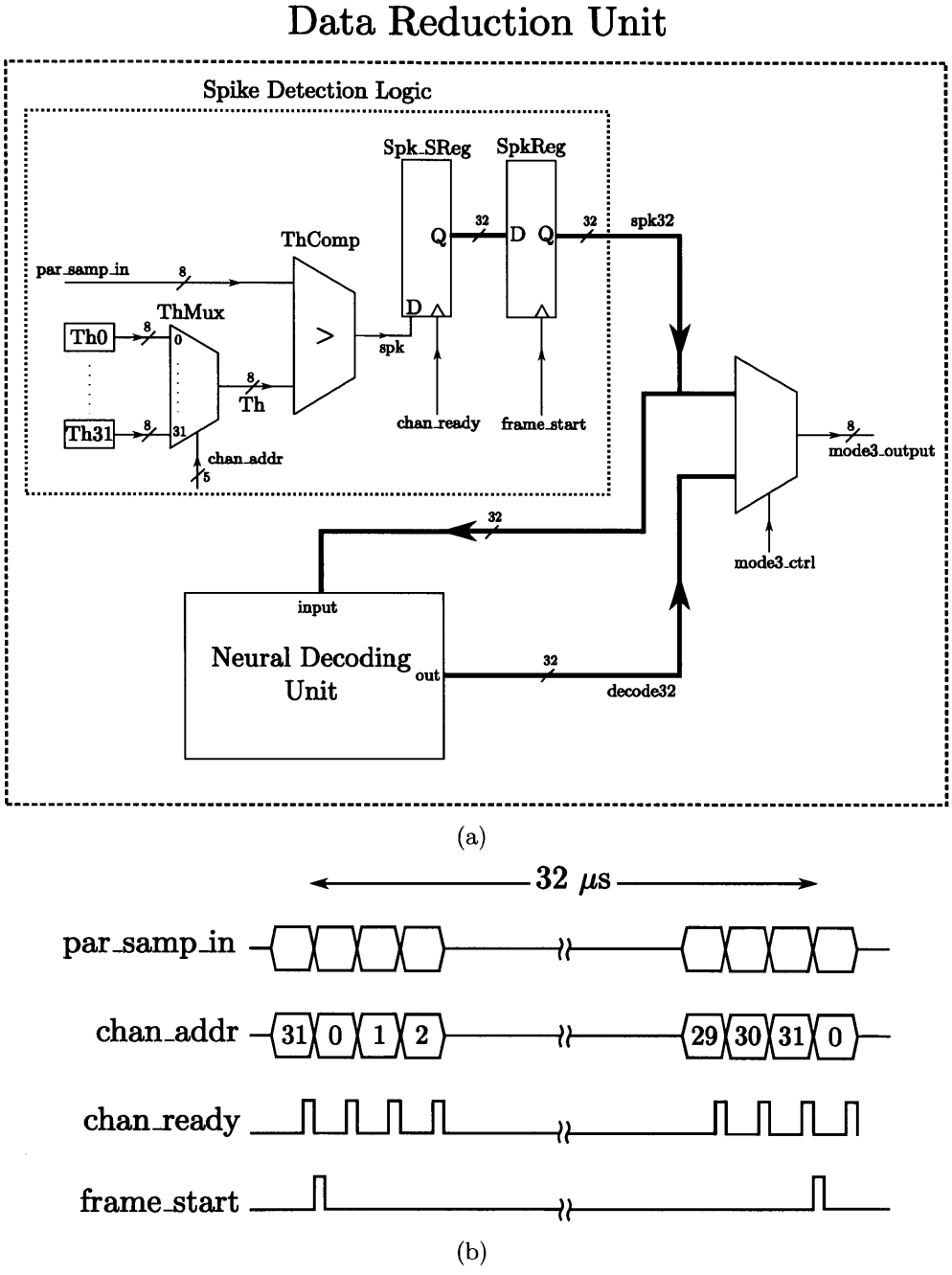
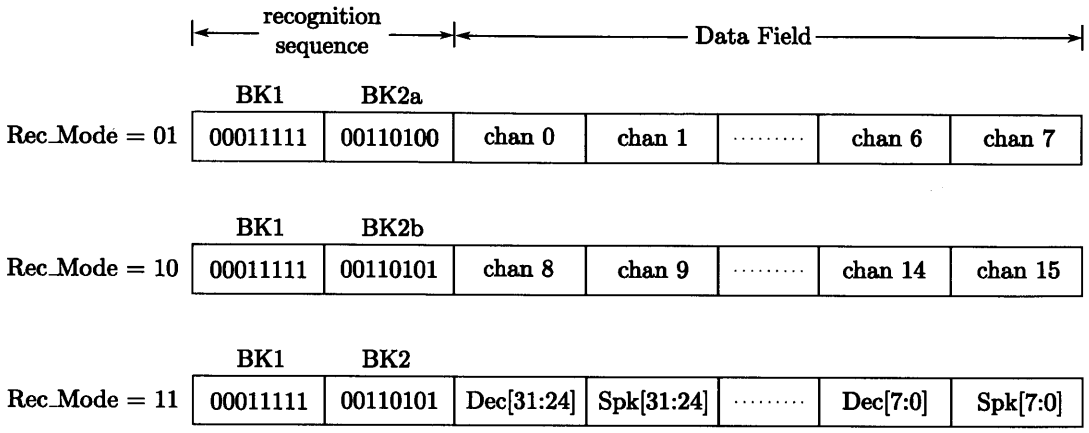


Figure 4-5: (a) Block diagram of the Data Reduction Unit. (b) Timing diagram of the input signals.

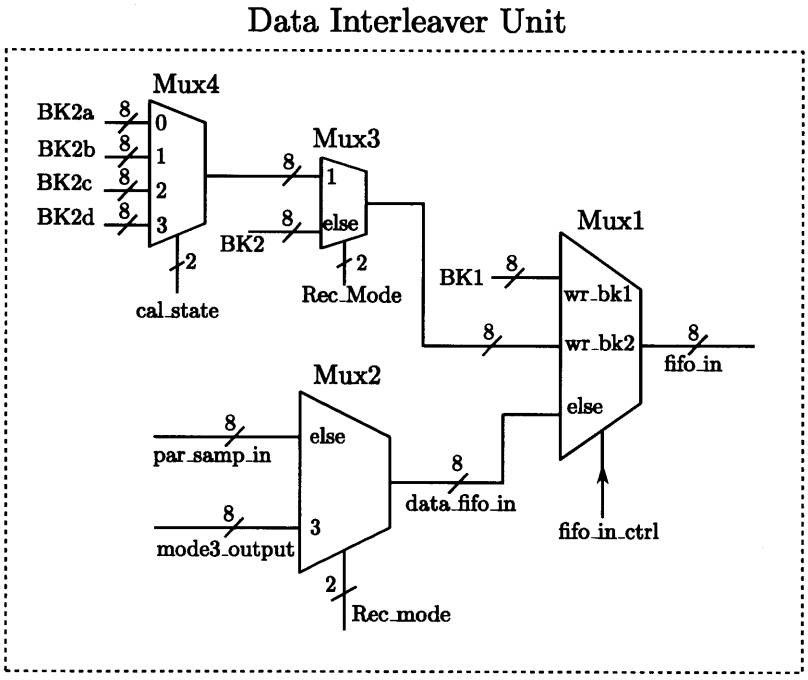
The spike detection thresholds of the 32 recording channels are stored in the 8-bit threshold storage registers Th0 to Th31. The contents of the threshold storage registers, Th0-Th31, must be programmed wirelessly through the downlink programming, prior to each recording session. During spike detection process, the threshold multiplexer, ThMux, which is controlled by chan_addr, sequentially multiplexes the thresholds stored in Th0 to Th31 to the input Th of the comparator ThComp. The incoming digitized neural data (par_samp_in) of the recording channel, which is specified by chan_addr, is then compared to its corresponding spike detection threshold on that channel to produce a one-bit output Spk. The output Spk is a 1 if the digitized data on par_samp_in is greater than the threshold, and is a 0 otherwise. Thus, for every 8-bit sample that comes in, the Spike Detection Logic reduces it to a 1-bit data, resulting in an 8:1 compression ratio. The output signal Spk is then shifted into a serial-in-parallel-out shift register Spk_SReg upon the positive edge of chan_ready. After threshold detection has been performed on all the 32 channels in a 32 μs data frame specified by frame_start, the 32-bit output of the shift register Spk_SReg is registered by a 32-bit storage register SpkReg upon the positive edge of frame_start. Registering the output of Spk_SReg is to guarantee that the threshold crossing information of the neural data from all 32 channels in the previous data frame is available for further processing in the next 32 μs data frame (on spk32 bus). The output of the Neural Decoding Unit is also registered by the signal frame_start, thus it is also unchanging during the next 32 μs data frame. The output of the Spike Detection Logic, spk32, and the output of the Neural Decoding Unit, decode32, are then multiplexed out to the output mode3_output to the Data Interleaver Unit.

4.4.2 Data Interleaver Unit

The main function of the Data Interleaver Unit is to organize the order of data that will be sent out to the Data Serializer Unit such that when the data is streamed out serially from the Data Serializer Unit, the serial output stream is in a format recognizable by the external unit, and thus can be synchronized to the data acquisition system and a PC. Depending on the content of the register Rec_Mode, the Data



(a)



(b)

Figure 4-6: (a) Output data format for different output modes. (b) Block diagram of the Data Interleaver Unit.

Interleaver Unit organizes the order of data into one of the three formats shown in Fig 4-6(a). The first two bytes of each data package form a recognition sequence, while the last eight bytes are reserved for the neural information that will be sent out to the external unit. In the *Calibration* mode (Rec_Mode = "01" in binary basis), the last two bits of the second byte of the recognition sequence are used to provide the information regarding the set of eight channels that are being streamed out. This information is important for the external unit because the set of eight channels that are being sent out is not fixed, but changed every two seconds. As shown in Fig. 4-6(a), when the second byte of the recognition sequence is "00110100" (BK2a), the data field contains the neural data from channel 0 to channel 7. When the second byte of the recognition sequence is "00110101" (BK2b), the data field contains the neural data from channel 8 to channel 15, and so on. In the *Stream8* and *Decode* mode (Rec_Mode = "10" and Rec_Mode = "11" in binary basis respectively), the recognition sequence is always "0001111" (BK1) followed by "00110101" (BK2) since the neural information in the data field is always from a know source. Therefore, we do not need to put additional information in the recognition sequence to tell what kind of neural information is inside the data field.

Figure 4-6(b) shows the block diagram of the Data Interleaver Unit. The multiplexer Mux1 controls the order of data to be sent to the Data Serializer Unit in a given 32 μs data frame (marked by frame_start in Fig. 4-5(b)). The operation of Mux1 is controlled by the control signal fifo_in_ctrl, which is generated from the Central Control Unit. The two bytes that comprise the recognition sequence are sent to the Data Serializer Unit at the beginning of the 32 μs data frame. After the bytes that comprise the recognition sequence have been sent, Mux1 only sends the neural information on data_fifo_in to the Data Serializer Unit, which can be from either the digitized neural data (par_samp_in) or the output of the Neural Decoding Unit (mode3_output) depending on the value of Rec_Mode. Choosing which value to write into the second byte of the recognition sequence is accomplished through the multiplexer Mux3 and Mux4. If the internal unit is not in the *Calibration* mode (Rec_Mode $\neq$ "01" in binary basis), Mux3 always choose "00110101" (BK2) as the

second byte of the recognition sequence. However, if the internal unit is in the *Calibration* mode, Mux3 will choose an 8-bit value from the set {BK2a,...,BK2d}, which is chosen by the multiplexer Mux4, to provide the external unit with the information regarding the content of the data field that is being sent out. The control signal for Mux4, cal_state, consists of the two most-significant bits of an 18-bit counter, which is clocked by the signal frame_start of Fig. 4-5(b). Since the period of frame_start is 32 μs, one counting period of the 18-bit counter is $2^{18} \times 32\ \mu s = 8.38$ s. Therefore, each set of eight channels will be streamed out for a duration of 8.38/4=2.1 s, before the Neural Data Processor changes to stream out the next set of eight channels.

4.4.3 Data Serializer Unit

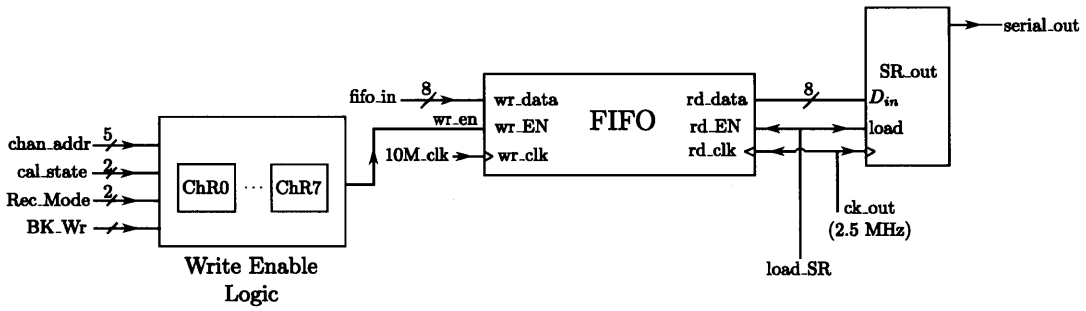


Figure 4-7: Block diagram of the Data Serializer Unit.

To communicate with the external unit, the internal unit must transmit the neural information at a uniform data rate. However, since the internal unit needs to be able to arbitrarily select eight recording channels to be sent out, it needs some kinds of data buffering before serialization to guarantee a uniform output data rate. The Data Serializer Unit simply serves this purpose. The block diagram of the Data Serializer is shown in Fig. 4-7. Once a recording session starts, the data to be sent out are written into the FIFO at the write port (wr_data). The writing into the FIFO is controlled by the write enable signal wr_en, which is generated from the Write Enable Logic block. The signal wr_en is always asserted when BK_wr goes high, which indicates that the data on the fifo_in port is one of the bytes of the recognition sequence. The Write Enable Logic block contains eight 5-bit registers, ChR0-ChR7, that store the

addresses of the channels chosen to be sent out. The signal wr_en is asserted when the address on the chan_addr port matches the content of one of the registers ChR0-ChR7. In the *Calibration* mode, the contents of ChR0-ChR7 are cycled by the signal cal_state. If cal_state = "00" in binary basis, ChR0-ChR7 store the channel addresses 0-7 respectively. When cal_state = "01" in binary basis, ChR0-ChR7 store the channel addresses 8-15 respectively, and so on. In the *Stream8* mode, the contents of ChR0-ChR7 must be programmed by the user which indicate the recording channels to be sent out. In the *Decode* mode, writing the data from the Data Reduction Unit is much simpler because the 32-bit outputs of the Spike Detection Logic and the Neural Decoder in Fig. 4-5(a) are fixed for the whole 32 μs sampling period. As a result, we simply stream mode3_output into the FIFO the same way as we want to write neural data from channel 0 to channel 7 into the FIFO.

After 32 μs has passed since the start of a recording session, at least 10 bytes of data will already be written into the FIFO. We can then stream the data from the FIFO at the read port (rd_data) into a parallel-in-serial-out shift register SR_out. The signal load_SR, generated from the Central Control Unit, acts as the read enable signal for the FIFO and, at the same time, acts as the load enable signal for the shift register SR_out. The output clock, clk_out, is generated from the Central Control Unit and has a frequency of 2.5 MHz, which is equal to the desired output data rate of the system. The clock clk_out is then used as a read clock for the FIFO and the clock of the shift register SR_out. The signal load_SR is asserted every eight cycles of clk_out which corresponds to the time it takes for SR_out to serially shift out an 8-bit byte of neural data from the FIFO.

4.4.4 Central Control Unit

The Central Control Unit is responsible for generating the control signals for the Neural Data Processor. The block diagram of the Central Control Unit is shown in Fig. 4-8(a). The main building block of the Central Control Unit is a 9-bit up-counter C320. The counter is clocked by the 10 MHz clock and repeatedly counts from 0 to 319 to divide the 32 μs data period into 320 equally time segments. This results in

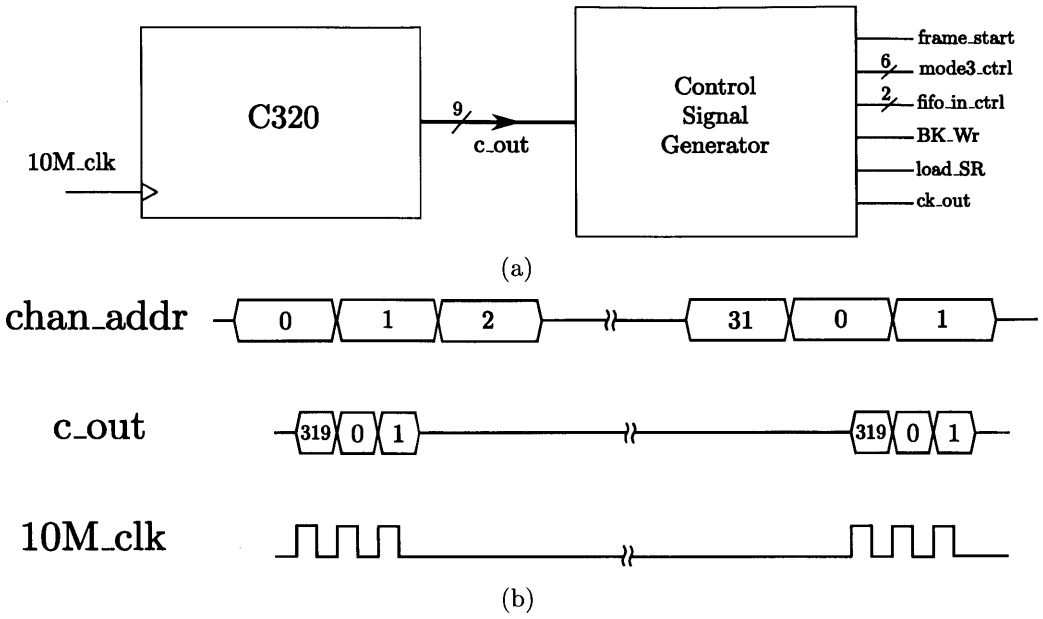


Figure 4-8: (a) Block diagram of the Central Control Unit. (b) Timing diagram of the 9-bit counter C320 used in the Central Control Unit.

a 100 ns-period per each time segment, which is the finest time scale in our system. Note that this strategy was once employed in the design of the Digital Control Unit in Chapter 3. The counter is synchronized to the channel address information on chan_addr such that it starts at 0 right when chan_addr = 1, and finishes counting 319 (in decimal basis) just exactly at the same time when chan_addr is changing from 0 to 1. The timing diagram of the counter C320 is shown in Fig. 4-8(b). The 9-bit output of the counter C320 is then used by the Control Signal Generator logic block to generate the required control signals for the Data Reduction Unit, the Data Interleaver Unit, and the Data Serializer Unit in the same fashion as in Section 3.6 of Chapter 3.

4.5 Operation Control Unit

As the name implies, the Operation Control Unit is responsible for controlling the operation of the internal unit. The main processing block of the Operation Control Unit is a synchronous state machine whose states determine the settings of the internal

unit.

4.5.1 Design Considerations of the Operation Control Unit

Before we discuss the design of the Operation Control Unit and the required inputs and outputs of the state machine, let's consider the following design considerations.

Transmit or Receive

The data telemetry system in [34] is a half-duplex system, meaning that the system can send the data in only one direction at any given time. When the system is in the uplink mode (data flows from the internal unit to the external unit), it does not alow the user to send commands to reconfigure the internal unit. Similarly, when the system is in the downlink mode (data flows from the external unit to the internal unit), the system cannot stream out neural information from the internal unit to the external unit. To configure the data telemetry system into either of the transmission modes, the user needs to set control switches on both the internal and the external data telemetry units accordingly. Let's denote these switches as Tx_Rx switches for both the internal and the external data telemetry unit. For the uplink mode, the Tx_Rx of the internal data telemetry unit must be set to 1 (meaning "transmit"), while the Tx_Rx of the external data telemetry unit must be set to 0 (meaning "receive"). For the downlink mode, the opposite switch configurations apply for both the internal and external data telemetry units. Setting the Tx_Rx switch of the external unit is trivial, since the user has an easy access to this switch at all time. However, the user has no direct access to the Tx_Rx switch on the internal unit, unless it is achieved through some wireless means. Obviously, we cannot configure the setting of the internal Tx_Rx switch through the same wireless data telemetry system, because once the internal data telemetry system is in the "transmit" mode, it cannot receive any downlink data. A different overriding scheme needs to be employed. For this purpose, we use magnetic field to control whether the internal data telemetry system should be in the "transmit" or the "receive" mode. An inverter circuit, consisting of of

a magnetic switch [2] and a pulled-up resistor, has been included on the internal unit to generate a signal HS as an input into the state machine. The HS signal is used to determine whether the Tx_Rx switch of the internal data telemetry unit should be set for the "transmit" or the "receive" mode. During normal recording session, the magnetic field is not applied to the magnetic switch, and thus the magnetic switch is open and HS=0. Based on this value of HS, the FPGA sets the switch Tx_Rx of the internal data telemetry unit to 1, making it in the "transmit" mode. As a result, neural information can be transmitted from the internal unit to the external unit. However, when the user needs to send commands to the internal unit, he can apply magnetic field to the magnetic switch, causing the switch to close and HS=1. With this value of HS, the FPGA sets the Tx_Rx switch of the internal data telemetry unit to 0, causing it to be in the "receive" mode, and thus ready to receive commands from the user.

Battery Charging or Recording

As mentioned earlier, we choose to power our internal unit from an implantable battery to achieve high-quality neural recordings. As a result, a battery charging IC [15] is included on the internal unit, which can be used to recharge the implantable battery as needed. During a battery charging session, the RF power is transferred from the external unit to the internal unit, where it is rectified to create a 5 V DC voltage on the internal unit. This DC voltage is then used by the battery charging IC to charge the implantable battery. It thus makes sense that battery charging should be done while a recording session is not in progress such that the RF power does not corrupt the ongoing neural recording session. To ensure that most of the current obtained from the rectified RF voltage is directed toward charging the battery, instead of being drained by other subsystems on the internal unit, it makes sense that other subsystems which are not required to operate during the battery charging session such as the neural recording IC, the internal data telemetry unit, the Neural Data Processor on the FPGA, and the 10 MHz crystal oscillator should be turned off. Therefore, the state machine needs an input to notify when the battery charging is in

session such that other subsystems can be shut down accordingly. For this purpose, we create a digital flag VR from the rectified DC voltage. Whenever the RF power from the external unit is enough to create the rectified DC voltage on the internal unit of more than 5 V, the flag VR is raised to 1. The signal VR is then used by the state machine to specify the charging status of the internal unit. Whenever $VR = 1$, the state machine shuts down other subsystems on the internal unit such that most of the output current from the battery charging IC goes into charging the battery instead of being drained by other running subsystems.

Recording or Idle

In order to maximize the time between recharges of the implantable battery, it is important to be able to force the internal unit into idle when it needs not operate. Thus, the state machine needs an input to specify whether the internal unit should be in a recording session, or should be in the idle state in which most of the electronics on the internal unit are shut down to save power. For this purpose, we create a storage register S on the FPGA whose content is used as an input to the state machine to specify whether the internal unit should be powered up or powered down. If $S = 1$, the internal unit should be powered up such that neural recording can be performed. If $S = 0$, the internal unit should be powered down to save power from the implantable battery. To set the content of the register S, the user must program it through the wireless data telemetry system. Note that S can be set in the same manner, during the downlink programming phase (through magnetic activation), as other parameters that are used in the Neural Data Processor.

Low Battery

To preserve battery health, it is advantageous to not deeply discharge the battery [33], [53]. Therefore, the internal unit should be able to notify when the battery is about to be deeply discharged such that it can shut down the rest of the system. To incorporate this feature, we have included a low battery flag LB as an input to the state machine. The low battery flag LB is provided by a commercially-available step

down switching regulator (LTC3620, Linear Technology Corp.) [1]. When the battery voltage falls below 3 V, the flag *LB* is raised to 1, indicating that the internal unit should go into idle to minimize current drawn from, and thus avoid damaging the battery.

4.5.2 Outputs of the Operation Control Unit

Depending on its current state, the state machine produces three outputs that are used for controlling the internal unit. Referring to Fig. 4-3, the first output is the Tx_Rx signal, which is used for controlling the Tx_Rx switch of the internal data telemetry system. The second output is the En_1.8V signal, which is used for turning on/off the 1.8 V supply domain. The third output is the LED signal, which is used for blinking the LED, to indicate the charging status of the internal unit. Since the 1.8 V supply domain powers the 32-channel neural recording IC and the 10 MHz crystal oscillator, the power drawn from this supply domain comprises the majority of the total power drawn from the implantable battery. Furthermore, once the 1.8 V supply domain is turned off, the 10 MHz clock is disabled causing the Neural Data Processor, even if it is powered from the 1.2 V supply domain, to stop operating and stop drawing current from the 1.2 V supply domain. Thus, the En_1.8V signal also indirectly acts as a clock gating signal [12, 13, 47] for the Neural Data Processor. Note that this strategy only eliminates the dynamic power dissipation of the Neural Data Processor, however, the leakage power still exists. The 1.2 V supply domain is always left operating since it also powers the state machine, which must continuously operate even when the internal unit is not in a recording session. Furthermore, the 2.5 V supply domain is always on since it needs to power the 32 kHz crystal oscillator which produces the clock signal for the state machine. However, the 32 kHz crystal oscillator only draws about 1 μA of current, which is considered negligible in our system. The 2.5 V supply domain also powers the LED such that when the battery charging is in progress, the LED blinks notifying the charging status.

4.5.3 Finite State Machine of the Operation Control Unit

Based on previous discussions, we divide the operation of the internal unit into four states as follows:

1. *Idle*: The 1.8 V supply domain is shut down, disabling the 32-channel neural recording IC, and the 10 MHz crystal oscillator. As a result, the Neural Data Processor on the FPGA is disabled, thus saving power drawn from the 1.2 V supply domain.

2. *Charge.* The 1.8 V supply domain is shut down such that most of the current from the rectified DC voltage is used for charging the battery. The on-board LED blinks, notifying the user that the implantable battery is being charged.

3. *Programming.* The Tx_Rx switch of the internal data telemetry unit is set to 0, causing it to be in the "receive" mode and ready to receive commands from the external unit. All the supply domains are powered up.

4. *Recording.* All the supply domains are powered up. The neural recording IC, the 10-MHz crystal oscillator, and thus the Neural Data Processor are operating. The Tx_Rx of the data telemetry unit is set to 1, causing it to be in the "transmit" mode and continuously transmitting neural information to the external unit.

The state diagram for controlling the state of operation of the internal unit is shown in Fig. 4-9. Table 4.2 summarizes the roles of the input signals into the state machine.

Table 4.2: Table summarizing the purposes of state machine's inputs

Input Signal	Action if asserted
HS	Go to *Programming* state
LB	Low battery (< 3 V), go to *Idle* state
VR	Received RF power, go to *Charging* state
S	Recording and transmitting neural information.

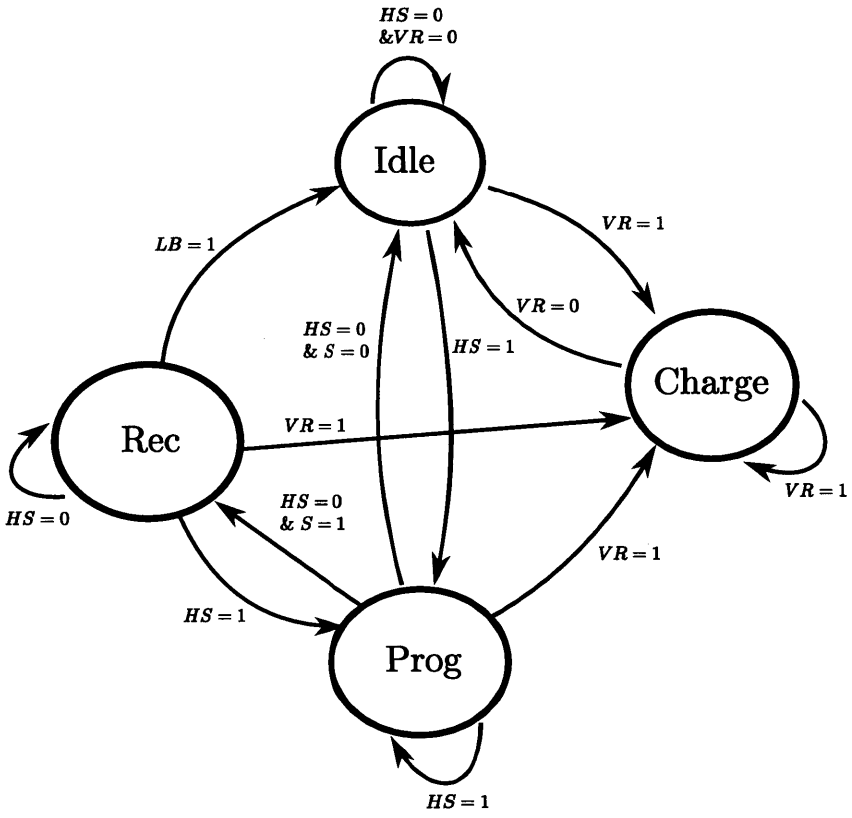


Figure 4-9: State diagram of the state machine in the Operation Control Unit.

4.6 Programming Interface Unit

The Programming Interface Unit of Fig. 4-3 has a structure similar to the serial programming interface of the 32-channel neural recording IC described in Section 3.5 of Chapter 3. It consists of a 56-bit shift registers that are clocked by the programming clock (prog_clk in Fig. 4-3) from the internal data telemetry unit. The required input data package is similar to that shown in Fig. 3-31. The data package consists of a 13-bit recognition sequence. However, the 8-bit address field is used as an 8-bit "Op Code" field. The "Op Code" field is used to specify which parameter registers on the FPGA to be programmed. The 35-bit payload field contains the data to be programmed into the parameter registers. Once the recognition sequence and the "Op Code" are detected by an internal logic block of the Programming Interface Unit, the data in the payload field are latched into the corresponding parameter registers addressed by the "Op Code". Once the parameter registers have been programmed, the programming data and programming clock pins (prog_dat and prog_clk in Fig. 4-3) are pulled down by the pull-down resisters attached to those pins. This is to prevent the inputs of the Programming Interface Unit from floating, which might cause an accidental reprogramming of the parameter registers. Therefore, the parameter registers will retain their contents until the power is cut from the internal unit, or until the user intentionally reprograms the parameters. Table 4.3 summarizes the roles of the parameter registers that must be programmed by the user.

Table 4.3: Parameter registers to be programmed by the user.

Registers	**Purpose**
Th0-Th31	storing 8-bit thresholds of all 32 channels
ChR0-ChR7	storing 5-bit addresses of eight channels to be streamed out in the *Stream8* mode
Rec_Mode	2-bit register specifying the output format of Neural Data Processor
S	specifying if the internal unit should be recording, or should be idle

Table 4.4 summarizes the values of "Op Code" for uses in programming different parameter registers of the internal unit.

Table 4.4: Table summarizing the purposes of different "Op Code".

Registers	Op Code	comment
Th0, Th8, Th16, Th24	0001 0001	payload: 8-bit thresh. values of 4 chan. in module 1
Th1, Th9, Th17, Th25	0001 0010	payload: 8-bit thresh. values of 4 chan. in module 2
Th2, Th10, Th18, Th26	0001 0011	payload: 8-bit thresh. values of 4 chan. in module 3
Th3, Th11, Th19, Th27	0001 0100	payload: 8-bit thresh. values of 4 chan. in module 4
Th4, Th12, Th20, Th28	0001 0101	payload: 8-bit thresh. values of 4 chan. in module 5
Th5, Th13, Th21, Th29	0001 0110	payload: 8-bit thresh. values of 4 chan. in module 6
Th6, Th14, Th22, Th30	0001 0111	payload: 8-bit thresh. values of 4 chan. in module 7
Th7, Th15, Th23, Th31	0001 1000	payload: 8-bit thresh. values of 4 chan. in module 8
ChR0-ChR3	0010 0001	payload: 5-bit 4 chan. addr. to be stored in ChR0-ChR3
ChR4-ChR7	0010 0010	payload: 5-bit 4 chan. addr. to be stored in ChR4-ChR7
Rec_Mode	0011 0001	Rec_Mode=01 (*Calibration*)
	0011 0010	Rec_Mode=10 (*Stream8*)
	0011 0011	Rec_Mode=11 ((Decode))
S	0100 0000	system goes into idle
	0100 0001	system is recording

4.7 Physical Design of the Internal Unit

The internal unit is designed for implantation in a non-human primate (a rhesus macaque). Figure 4-10 shows the original plan and the physical dimensions of the internal unit relative to the implantable area on a rhesus macaque's skull. The internal unit is intended to be implanted between the skull and the scalp and is integrated on a flexible PCB substrate such that it can be fitted to the curvature of the skull. The black dots in Fig. 4-10 represent the screw holes on the internal unit for tightening the unit to the skull. Note that the anterior part of the skull is toward the top of the figure (labeled with the letter "A"), while the posterior part is toward the bottom of the figure (labeled with the letter "P"). The part of the internal unit labeled "Flexible PCB" houses most of the circuit components of the internal unit, while the part labeled "Stacked Coils" is a support platform for holding the concentric data and power coils. The internal unit contains two nano strip connectors (A79041-01, Omnetic Connector Corp., Minneapolis, MN) for connecting two microelectrode arrays to the internal unit.

Figure 4-11(a) shows our assembled internal unit with some circuit components on the PCB labeled, while Fig. 4-11(b) shows just the flexible PCB substrate when it is flexed. The PCB contains 5 routing layers and one ground plane. The main com-

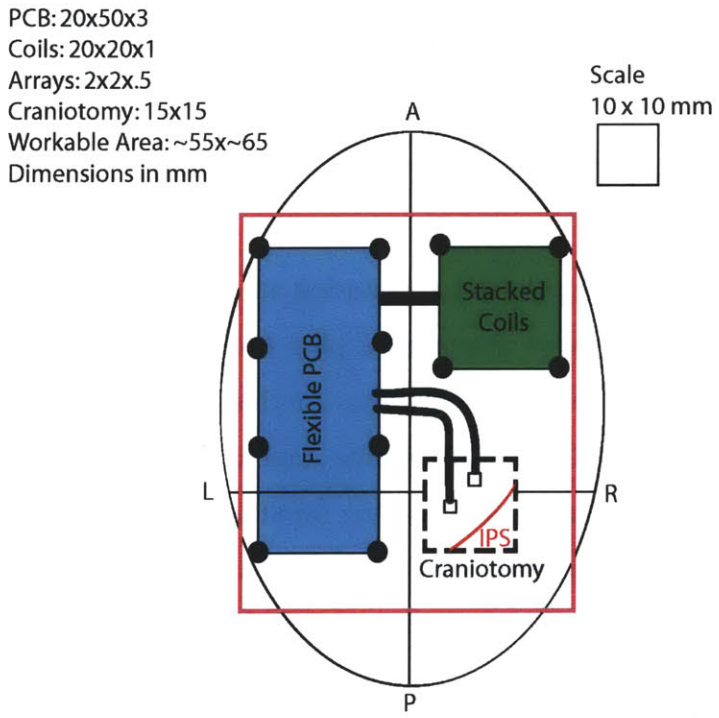
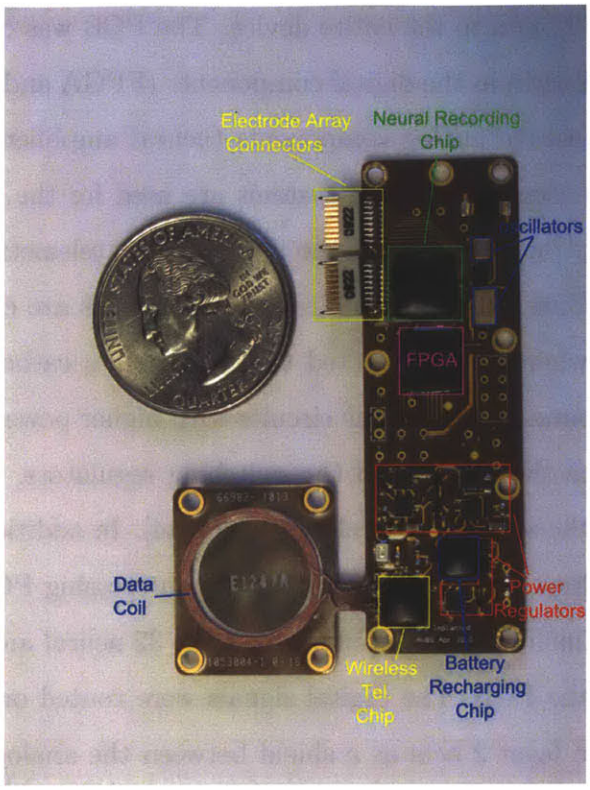
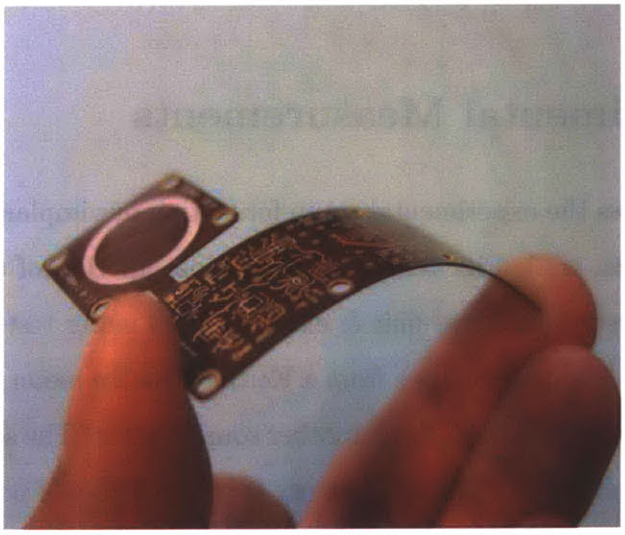


Figure 4-10: Conceptual diagram showing the planned physical dimensions of the internal unit relative to the surgery area on a rhesus macaque's skull.



(a)



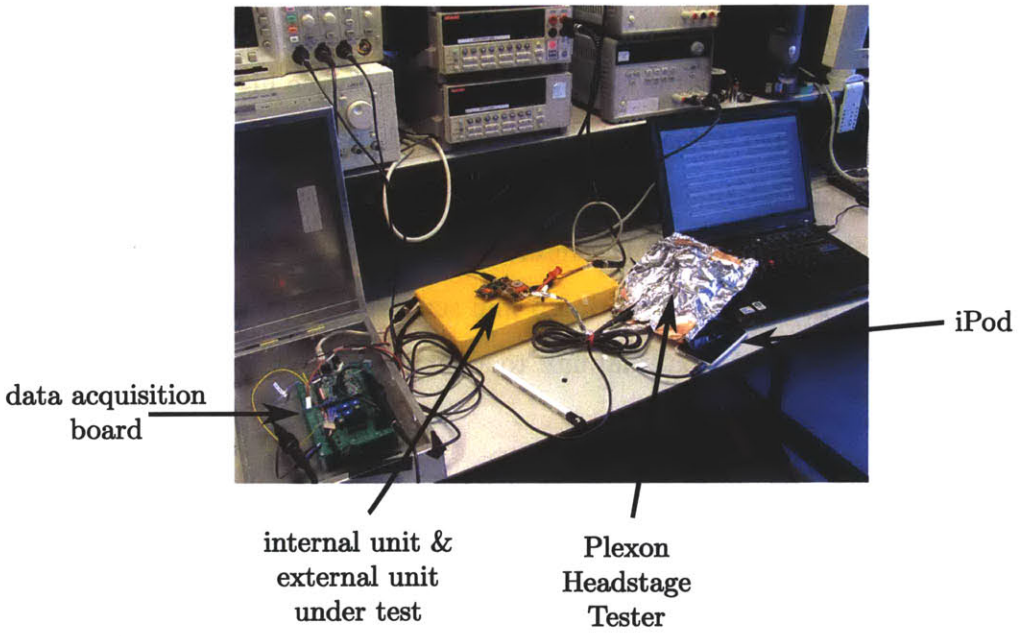
(b)

Figure 4-11: (a) Populated internal unit. (b) Flexible PCB substrate of the internal unit.

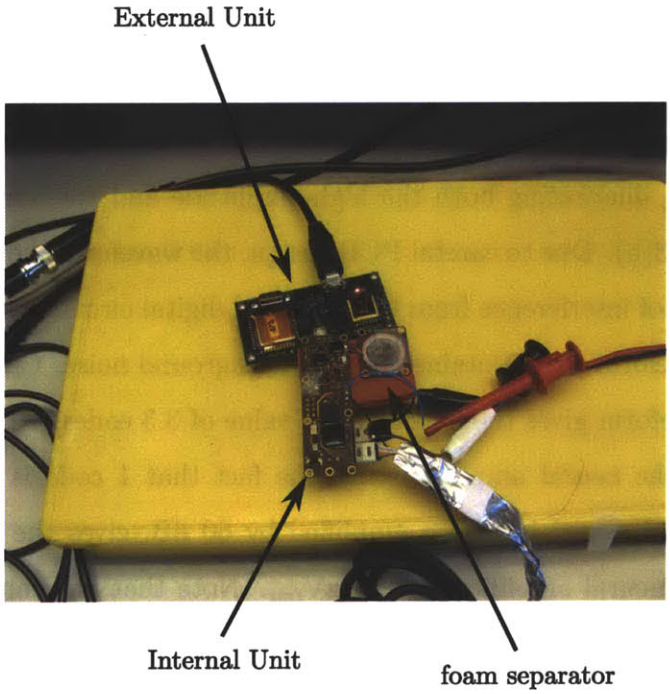
ponent of the PCB has dimensions of 18 mm × 56 mm, while the coil platform adds an area of 18 mm × 22 mm to the entire device. The PCB was carefully designed to minimize noise coupling from the digital components (FPGA and on-board switching regulators) to the sensitive analog components (neural amplifiers on the 32-channel neural recording IC). Separate ground systems are used for the Power Management & Battery Charging Unit, the FPGA, the internal data telemetry unit, and the 32-channel neural recording IC. The grounds of these circuits are combined at a single point on the PCB, where it is connected to the battery's cathode. This technique helps prevent the ground current of the circuits with higher power and higher switching activities, such as the FPGA and the switching regulators, from disturbing the ground potential of the sensitive neural amplifiers [44]. In addition, careful shielding was employed to reduce capacitive coupling between crossing PCB traces. The sensitive analog signals including the input wires of the 32 neural amplifiers were routed on the top layer of the PCB. The digital signals were routed on layers 3, 4, and 5. The ground plane on layer 2 acts as a shield between the analog signals on the top layer and the digital signals on layers 3 and below.

4.8 Experimental Measurements

Figure 4-12(a) shows the experimental setup for testing our implantable wireless neural recording system, while Fig. 4-12(b) shows a close-up view of the overall wireless neural recording system (internal unit & external unit) under test. The internal unit is powered by a 3.6 V supply voltage from a Keithley source meter, while the external unit is powered by a 3.3 V supply from another source meter. The source meter is used to power the internal unit instead of a battery for the purpose of power measurements during different states of operation of the internal unit. Neural signals are fed into one of the electrode array connectors (to 16 inputs of the internal unit), while another array connector is grounded. An audio file containing neural signals is played from an iPod. The audio output of the iPod is fed to a Plexon headstage tester [3] where the neural signals are attenuated by a factor of 1000 before being input into all the



(a)



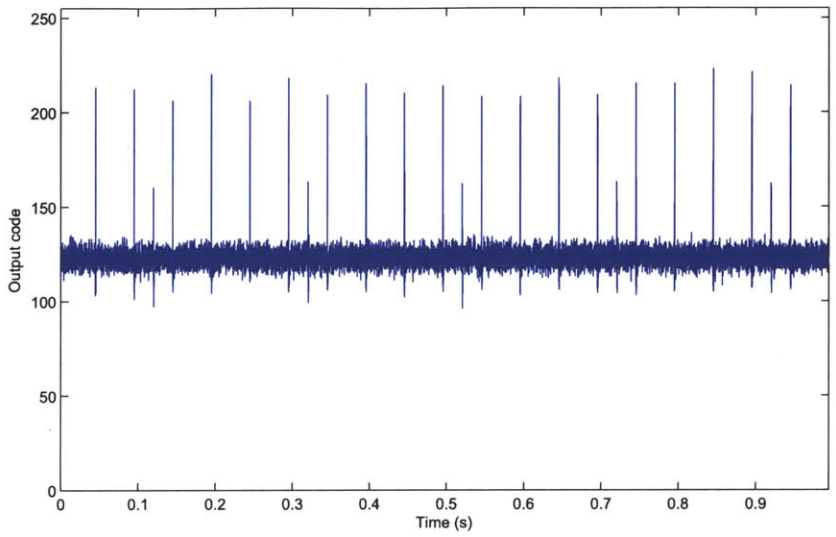
(b)

Figure 4-12: (a) Experimental setup of the implantable wireless neural recording system. (b) Internal unit and external unit under test.

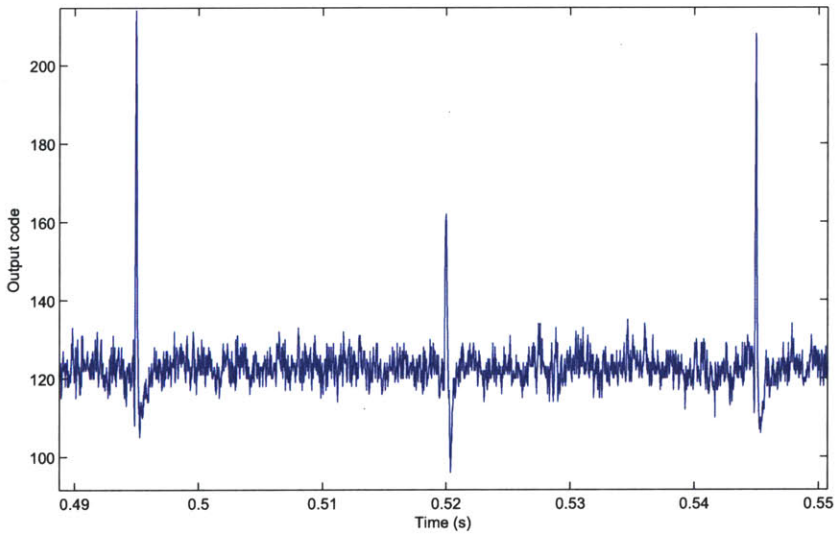
16 channels of the array connector in parallel. The headstage tester also mimics real recording conditions of real recording electrodes by providing high output impedance and adding thermal noise to the neural signals. Due to the high impedance on each channel of the headstage tester, the whole unit needs to be wrapped in an aluminum foil for shielding the high-impedance nodes against power line interference. As shown in Fig. 4-12(b), the data coil of the internal unit and that of the external unit are placed concentric to each other, with a 5-mm separation by a piece of foam. A USB cable is connected from the external unit to a data acquisition system where the data is streamed into a PC.

During a normal recording session (in the *Stream8* mode) when all the subsystems of the internal unit are operating and the data are being transmitted to the external unit, the internal unit consumes a total of 6.4 mW (1.8 mA from a 3.6 V supply). Note that the internal unit still works when the power supply is reduced to 3 V. However, we choose a supply voltage of 3.6 V since this is the output voltage of a Li-ion battery. Figure 4-13(a) shows the received neural data from the PC from one of the recording channels. The neural amplifier's gain for this recording is 60 dB. The close-up version of this waveform, illustrating both the high-amplitude and low-amplitude spikes, is shown in Fig. 4-13(b). Due to careful PCB design, the waveform shown in Fig 4-13(a) expresses no sign of interference from the on-board digital circuits. Figure 4-14 shows a part of the waveform that contains only the background noise. Calculating the rms value of this waveform gives the output noise value of 3.3 codes. Referring this value to the input of the neural amplifier, with the fact that 1 code is equal to $1/2^8$ V and that the input signal has been amplified by 60 dB, gives the noise referred to the input of the neural amplifier of 12.9 μV_{rms}. Note that the value of the resister used on the Plexon headstage tester for adding thermal noise to the neural signals is 500 kΩ. For the 12-kHz bandwidth of the neural amplifier, the thermal noise due to this resister is 12.5 μV_{rms}. Therefore, the noise seen on the output waveform of our wireless neural recording system is mostly from the headstage tester, and not from intrinsic noise of our system.

We have also tested different states of operation of the internal unit. The downlink



(a)



(b)

Figure 4-13: Digitized neural data from one of the recording channels (gain = 60 dB) of the implantable wireless neural recording system (a) Long-time trace. (b) Short-time trace.

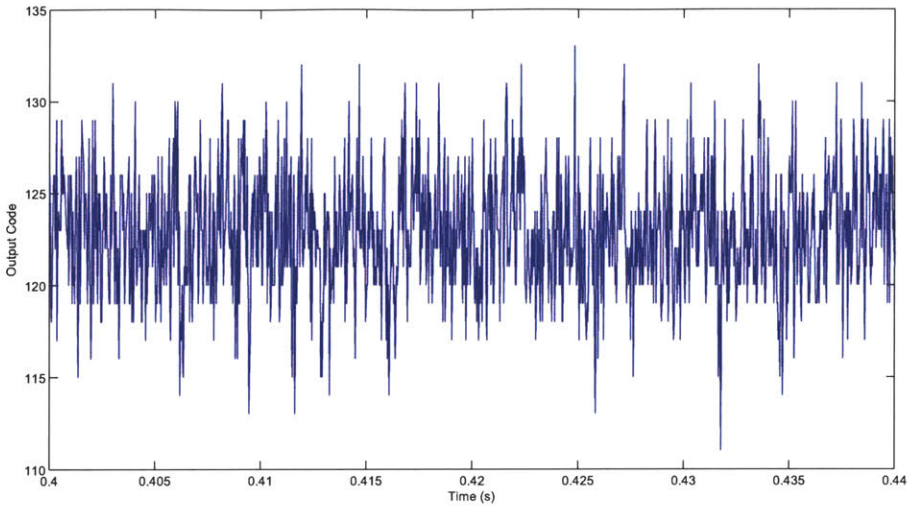


Figure 4-14: A short-time trace showing the noise of the waveform in Fig. 4-13(a).

programming works correctly. We are able to force the internal unit into the *Programming* mode by applying magnetic field to the on-board magnetic switch, and are able to configure all the parameters of the internal unit correctly. When the internal unit is programmed into the *Idle* mode, the 1.8 V supply domain is powered down as expected, and the internal unit draws a total current of 400 μA from the 3.6 V supply (from the Keithley source meter). This standby power is due to the standby current of the off-the-shelf switching power regulators and linear regulators in the power management & battery charging unit. Due to strong coupling between the data and power coils of the internal unit, the two coils cannot be placed concentric to each other during the battery charging phase. The strong coupling between the two internal coils causes the internal data telemetry unit, which is connected to the data coil, to leak significant amount of current (around 4 mA) from the 2.5 V supply domain during the battery charging phase. Thus, the output current from the battery charging IC is drained away due to the power leakage from the 2.5 V supply domain, and could not be directed toward charging the battery. However, this problem can be solved simply by separating the two coils such that they do not overlap. We test the battery charging circuitry by transferring RF power to the receiving internal power coil on the internal unit. Once the rectified voltage of more than 5 V is detected on

the internal unit, the internal unit goes into the *Charging* mode. During this mode, the current of 1.7 mA is delivered to the Keithley source meter, indicating that the battery charging IC is successfully delivering current to charge an implantable battery as designed.

Table 4.5: System Level Performance

Specification	***Value***
No. of Channels	32
Power Supply	3.6 V (battery)
Output Modes	i) waveform snippets of all 32 channels ii) 8 channels of raw waveform iii) decoded & thresholded spike information
Uplink Data Rate	2.5 Mbps
Downlink Data Rate	300 kbps
Amplifier:	
Gain	49-66 dB
Low cut-off	0.126 Hz or 350 Hz
High cut-off	293 Hz or 12 kHz
Input-referred noise	5.4-11.2 μV_{rms}
Power Consumption:	
Recording	6.4 mW
Stand-by	1.4 mW
Figure of Merit	80 pJ/(ch · bit)
Size:	
Main PCB	1.8 cm × 5.6 cm
Coil platform	1.8 cm × 2.2 cm

Table 4.5 summarizes the performance of our implantable wireless neural recording system. In order to compare the energy efficiency of various implantable wireless neural recording systems reported in the literature, we have devised a figure of merit (FOM) which bases on total power consumption, number of recording channels, and the output data rate of the system. The figure of merit can be calculated from FOM = P_{total}/(No. of channels × Output data rate). Our system achieves a figure of merit of 80 pJ/(Ch·bit), which places it among the most energy-efficient implantable wireless neural recording systems to date. Table 4.6 compares the performance of our neural recording systems with those published earlier in the literature. Note that we do not include the FOM of the work in [17] because it would not be a fair

Table 4.6: Comparison to other state-of-the-art implantable wireless neural recording systems

Reference	**[26]**	**[62]**	**[17]**	**This work**
No. of Channels	100	64	6	32
Supply Voltage	3.55 V	1.8 V	3 V	3.6 V
Up. Data Rate	330 kbps	2 Mbps	9.6 kbps	2.5 Mbps
Down. Data Rate	6.5 kbps	2 Mbps	NA	300 kbps
Tx. Range	a few cm	a few cm	9 m	a few cm
Amplifier: gain	60 dB	60 dB	46 dB	49-66 dB
Low cut-off	300 Hz	<10 Hz-100 Hz	a few Hz	0.126 Hz, 350 Hz
High cut-off	5 kHz	9.1 kHz	1 kHz	293 Hz, 12 kHz
Power	13.5 mW	14.4 mW	66 mW	6.4 mW
FOM (pJ/(Ch·bit))	450	112	NA	80
Sizes	NA	1.4 cm × 1.55 cm	2.5 cm × 2.5 cm	1.8 cm × 5.6 cm

comparison. The system in [17] was designed to transmit to a long distance of 9 m, and thus its wireless telemetry system consumes much more power than other systems in Table 4.6.

4.9 Conclusion

In this chapter, I have presented the design of an implantable wireless neural recording system. The design of the overall system including the internal unit and the external unit have been presented, while the emphasis is on the neural signal processing aspect on the internal unit. The implantable unit (internal unit) consists of the 32-channel neural recording IC which is the topic of Chapter 3, the power management & battery charging unit, the on-board low-power FPGA, and the impedance-modulation wireless telemetry system. The internal units utilizes separate data coil and power coil for data and power transfer. The system can amplify and digitize neural signals from 32 recording electrodes. The digitized neural signals are processed by the on-board FPGA to reduce the amount of data that must be transmitted to the external world. The system was successfully tested on the lab bench and exhibited high-quality recordings. By combining state-of-the-art ASICs, commercially-available FPGA and discrete components, the system achieved excellent energy efficiency, while still offering design flexibility during the system development phase. The internal unit's power consumption of 6.4 mW from a 3.6 V supply and the wireless output data rate of 2.5 Mbps place it among the most energy-efficient implantable wireless neural recording systems reported to date.

Chapter 5

Conclusions

5.1 Summary

This thesis focused on the development of an implantable wireless neural recording system. The major design goal is to minimize the total power consumption of the implantable unit to avoid excessive heat dissipation in the area of implantation and to maximize the lifetime of the implantable battery that powers the internal unit. The following contributions were made to this area:

- The design of an ultra-low-power neural recording amplifier was reported. The design utilized the folded-cascode amplifier topology with extreme current scaling technique and source-degenerated current mirrors to simultaneously minimize the input-referred noise of the amplifier and its overall power consumption. With such techniques, the amplifier achieved an input-referred noise close to theoretical limit of two transistors in an input differential pair, while most of its supply current was consumed only in this input differential pair.

- The design of an ultra-low-power 32-channel neural recording system (IC) was reported. By utilizing an energy-and-area-efficient amplifier, analog multiplexer, and digital-to-analog converter, the neural recording IC achieved very low power and small silicon area per recording channel.

- An adaptive biasing strategy was presented and utilized in designing the neural

amplifier. Combined with the system programmability, the adaptive biasing technique helps minimize the total power consumption of the overall IC, while still making it useful for different recording conditions.

These innovations were incorporated into the design of a fully-implantable wireless neural recording system. The implantable wireless neural recording system can amplify and digitize neural signals from 32 recording electrodes, and transmit neural information to the external unit at a rate of 2.5 Mbps. The system also incorporated an energy-efficient battery charging circuitry for charging an implantable battery that powers the internal unit. By incorporating a low-power FPGA and state-of-the-art ASICs, the implantable system achieved very low-power consumption, while offering design flexibility during the system development phase.

5.2 Future Work

Several areas are worth exploring to make the current neural recording system more useful in practical BMIs. These areas include:

- A clear disadvantage of the presented system is its size. To make the system useful in clinical applications, the size of the implantable unit should be reduced to approximately no larger than 2 cm $\times$ 2 cm. The size limitation of this work arises from the low level of system integration at the integrated circuit level (a tradeoff for design flexibility). By combining the subsystems described in this thesis into a single IC, the size of the implantable unit will be drastically reduced. This is one of the reasons we use an on-board FPGA, instead of a microcontroller, to implement the signal processing functions of the system. The hardware description language (HDL) for synthesizing the digital system on the FPGA can be conveniently ported to implement the same functions in a custom ASIC environment.

- A lot of improvements can be done toward reducing the total power consumption of the internal unit. A large fraction of the total power consumption of the

internal unit is in the 10 MHz clock generation process, chip-to-chip communications, and signal processing on the FPGA. Note that the power consumption of the 10 MHz crystal oscillator alone is about 1.5 mW. This power can be reduced to less than 100 μW by implementing the active circuitry part of such oscillator in the same technology in which the 32-channel neural recording IC was implemented. The power consumption of the FPGA amounts to about 2 mW. This same power can be reduced to less than 100 μW if the same digital functions are implemented in an ultra-low-power digital ASIC manner.

- For high performance BMIs, a high-channel-count system is necessary. The presented system was designed with only 32 input channels as a proof of concept. However, increasing the number of recording channels is an improvement worth pursuing. The designs of the circuit components in the 32-channel neural recording IC are suitable for scaling to a high-channel-count system. Note that the reason we chose to implement only 32 recording channels was not because of the die area limitation, but because of the system integration issues. Interfacing multielectrode arrays with a high-channel-count neural recording IC will need a more sophisticated integration technology than what we used in this thesis. A high-channel-count neural recording IC will require a larger number of I/O pads, which will make the pad pitch even smaller and the routing to these pads more difficult. In our system, routing 32 traces on a flexible PCB from the array connectors to the 32 input pads of the neural recording IC poses some design challenges due to limitations of the PCB technology. A more promising method is to use a silicon platform [62] in which the signals can be easily routed with ultra-fine traces in the same manner as in VLSI design.

Bibliography

[1] "LTC3620 Ultralow Power 15mA Synchronous Step-Down Switching Regulator". data sheet. http://cds.linear.com/docs/Datasheet/3620fa.pdf.

[2] "MMS series Reed Sensor". data sheet. http://www.meder.com/fileadmin/meder/pdf/en/Products/Reed_Sensors/Reed_Sensor_MMS_E.pdf.

[3] "Plexon Headstage Tester Unit Guide". data sheet. http://www.plexon.com/assets/pdf/HeadstageTesterUnitGuide.pdf.

[4] "IGLOO Low Power Flash FPGAs". data sheet, Nov 2009. http://www.actel.com/documents/IGLOO_DS.pdf.

[5] "CC8520 2.4 Ghz RF SoC for Wireless Digital Audio Streaming - PurePath Wireless". TI data sheet, Oct 2010. http://focus.ti.com/lit/ds/symlink/cc8520.pdf.

[6] R. A. Andersen, S. Musallam, and B. Pesaran. Selecting the signals for a brain-machine interface. *Current Opinion in Neurobiology*, 14(6):720–726, Dec 2004.

[7] M. W. Baker and R. Sarpeshkar. Feedback Analysis and Design of RF Power Links for Low-Power Bionic Systems. *IEEE Trans on Biomed. Circuits and Systems*, 1(1):28–38, Mar 2007.

[8] R. J. Baker. *CMOS Mixed-Signal Circuit Design.* John Wiley & Sons, Inc., 2009.

[9] T. Borghi, R. Gusmeroli, A. S. Spinelli, and G. Baranauskas. A simple method for efficient spike detection in multiunit recordings. *J. of Neuroscience Methods*, 163(1):176–180, Jun 15 2007.

[10] M. S. Chae, Z. Yang, M. R. Yuce, L. Hoang, and W. Liu. A 128-Channel 6 mW Wireless Neural Recording IC With Spike Feature Extraction and UWB Transmitter. *IEEE Trans. on Neural Systems and Rehab. Engineering*, 17(4):312–321, Aug 2009.

[11] Moosung Chae, Wentai Liu, Zhi Yang, Tungchien Chen, Jungsuk Kim, M. Sivaprakasam, and M. Yuce. A 128-channel 6mw wireless neural recording ic with on-the-fly spike sorting and uwb tansmitter. In *ISSCC 2008 Dig. Tech. Papers.*, pages 146 –603, Feb. 2008.

[12] A. P. Chandrakasa, S. Sheng, and R. W. Brodersen. Low-Power CMOS Digital Design. *IEEE J. of Solid-State Circuits*, 27(4):473–484, Apr 1992.

[13] A. P. Chandrakasan and R. W. Brodersen. Minimizing Power-Consumption in Digital CMOS Circuits. *Proceedings of the IEEE*, 83(4):498–523, Apr 1995.

[14] T. Delbruck and C.A. Mead. Adaptive photoreceptor with wide dynamic range. In *IEEE Inter. Symposium on Circuits and Systems, 1994*, volume 4, pages 339 –342 vol.4, May-2 Jun 1994.

[15] B. Do Valle, C. T. Wentz, and R. Sarpeshkar. An area and power-efficient analog li-ion battery charger circuit. *IEEE Trans. on Biomedical Circuits and Systems*, PP(99):1, 2011.

[16] C.C. Enz, F. Krummenacher, and E.A. Vittoz. An analytical mos transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications. *Analog Integrat. Circuits Signal Process*, 8:83–144, 1995.

[17] S. Farshchi, P. H. Nuyujukian, A. Pesterev, I. Mody, and J. W. Judy. A TinyOS-enabled MICA2-based wireless neural interface. *IEEE Trans. on Biomedical Engineering*, 53(7):1416 –1424, Jul 2006.

[18] Maysam Ghovanloo and Suresh Atluri. A wide-band power-efficient inductive wireless link for implantable microelectronic devices using multiple-carriers.

IEEE Trans. on Circuits and Systems I-Regular Papers, 54(10):2211–2221, Oct 2007.

[19] B. P. Ginsburg and A. P. Chandrakasan. An energy-efficient charge recycling approach for a SAR converter with capacitive DAC. pages 184–187. IEEE Inter. Symp. on Circuits and Systems (ISCAS), Kobe, Japan, May 23-26, 2005.

[20] B. Gosselin, A. E. Ayoub, J-F Roy, M. Sawan, F. Lepore, A. Chaudhuri, and D. Guitton. A Mixed-Signal Multichip Neural Recording Interface With Bandwidth Reduction. *IEEE Trans. on Biomed. Circuits and Systems*, 3(3):129–141, Jun 2009.

[21] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits.* John Wiley & Sons, Inc., 4th edition, 2001.

[22] K. S. Guillory and R. A. Normann. A 100-channel system for real time detection and storage of extracellular spike waveforms. *J. of Neuroscience Methods*, 91(1-2):21–29, Sep 1999.

[23] K.S. Guillory and R. A. Normann. A 100-channel system for real time detection and storage of extracellular spike waveforms. *J. of Neuroscience Methods*, 91:21–29, 1999.

[24] R. R. Harrison and C. Charles. A low-power low-noise CMOS amplifier for neural recording applications. *IEEE J. of Solid-State Circuits*, 38(6):958–965, Jun 2003.

[25] R. R. Harrison, R. J. Kier, C. A. Chestek, V. Gilja, P. Nuyujukian, S. Ryu, B. Greger, F. Solzbacher, and K. V. Shenoy. Wireless Neural Recording With Single Low-Power Integrated Circuit. *IEEE Trans. on Neural Systems and Rehab. Engineering*, 17(4):322–329, Aug 2009.

[26] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher. A low-power integrated circuit for a wireless 100-electrode neural recording system. *IEEE J. of Solid-State Circuits*, 42(1):123–133, Jan 2007.

[27] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, Jul. 13 2006.

[28] D. A. Johns and K. Martin. *Analog Integrated Circuit Design.* John Wiley & Sons, Inc., 1997.

[29] D. R. Kipke, R. J. Vetter, J. C. Williams, and J. F. Hetke. Silicon-substrate intracortical microelectrode arrays for long-term recording of neuronal spike activity in cerebral cortex. *IEEE Trans. on Neural Systems and Rehab. Eng.*, 11(2):151–155, Jun. 2003.

[30] B. Lenaerts and R. Puers. An inductive power link for a wireless endoscope. *Biosensors & Bioelectronics*, 22(7):1390–1395, Feb 15 2007.

[31] K. N. Leung and P.K.T. Mok. A sub-1-v 15-ppm/ deg;c cmos bandgap voltage reference without requiring low threshold voltage device. *IEEE J. of Solid-State Circuits*, 37(4):526–530, Apr 2002.

[32] H. Li, W. Zhao, and Y. Zhang. Micropower fully integrated CMOS readout interface for neural recording application. *Microelectronics Reliability*, 50(2):273–281, Feb 2010.

[33] D. Linden and T. B. Reddy. *Handbook of Batteries.* McGraw-Hill, 3rd edition, 2002.

[34] S. Mandal and R. Sarpeshkar. Power-Efficient Impedance-Modulation Wireless Data Links for Biomedical Implants. *IEEE Trans. on Biomed. Circuits and Systems*, 2(4):301–315, Dec 2008.

[35] E. M. Maynard, C. T. Nordhausen, and R. A. Normann. The Utah Intracortical Electrode Array: A recording structure for potential brain-computer interfaces. *Electroencep. and Clinical Neurophysiology*, 102(3):228–239, Mar. 1997.

[36] P. Mohseni and K. Najafi. A fully integrated neural recording amplifier with dc input stabilization. *IEEE Trans. on Biomedical Engineering*, 51(5):832–837, May 2004.

[37] A. F. Molisch. Ultra-Wide-Band Propagation Channels. *Proceedings of The IEEE*, 97(2, Sp. Iss. SI):353–371, Feb 2009.

[38] M. Mollazadeh, K. Murari, G. Cauwenberghs, and N. Thakor. Micropower CMOS Integrated Low-Noise Amplification, Filtering, and Digitization of Multimodal Neuropotentials. *IEEE Trans. on Biomed. Circuits and Systems*, 3(1):1–10, Feb 2009.

[39] S. Narasimhan, Y. Zhou, H. J. Chiel, and S. Bhunia. Low-power VLSI architecture for neural data compression using vocabulary-based approach. pages 134–137. IEEE Biomedical Circuits and Systems Conference, Montreal, Canada, Nov 27-30, 2007.

[40] Z. Nenadic and J.W. Burdick. Spike detection using the continuous wavelet transform. *IEEE Trans. on Biomed. Engineering*, 52(1):74 –87, Jan 2005.

[41] R. H. Olsson, D. L. Buhl, A. M. Sirota, G. Buzsaki, and K. D. Wise. Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays. *IEEE Trans. on Biomedical Engineering*, 52(7):1303–1311, July 2005.

[42] R. H. Olsson and K. D. Wise. A three-dimensional neural recording microsystem with implantable data compression circuitry. *IEEE J. of Solid-State Circuits*, 40(12):2796–2804, Dec 2005.

[43] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing.* Prentice Hall, 2nd edition, 1999.

[44] H. W. Ott. *Noise Reduction Techniques in Electronic Systems.* John Wiley & Sons, Inc., 2nd edition, 1988.

[45] Y. Perelman and R. Ginosar. An integrated system for multichannel neuronal recording with spike/LFP separation, integrated A/D conversion and threshold detection. *IEEE Trans. on Biomed. Engineering*, 54(1):130–137, Jan 2007.

[46] R. Puers and G. Vandevoorde. Recent progress on transcutaneous energy transfer for total artificial heart systems. *Artificial Organs*, 25(5):400–405, May 2001.

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

[48] B. Rapoport. *Glucose-Powered Neuroelectronics.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, April 2011.

[49] M. Rizk, C. A. Bossetti, T. A. Jochum, S. H. Callender, M. A. L. Nicolelis, D. A. Turner, and P. D. Wolf. A fully implantable 96-channel neural data acquisition system. *J. of Neural Engineering*, 6(2), Apr 2009.

[50] M. Rizk, I. Obeid, S. H. Callender, and P. D. Wolf. A single-chip signal processing and telemetry engine for an implantable 96-channel neural data acquisition system. *J. of Neural Engineering*, 4(3):309–321, Sept 2007.

[51] J. K. Roberge. *Operational Amplifiers: Theory and Practice.* John Wiley & Sons, Inc., 1975.

[52] G. Santhanam, M. D. Linderman, V. Gija, A. Afshar, S. I. Ryu, T. H. Meng, and K. V. Shenoy. HermesB: A continuous neural recording system for freely behaving primates. *IEEE Trans. on Biomedical Engineering*, 54(11):2037–2050, Nov 2007.

[53] R. Sarpeshkar. *Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems.* Cambridge University Press, 2010.

[54] R. Sarpeshkar, R. F. Lyon, and C. Mead. A low-power wide-linear-range transconductance amplifier. *Analog Integrated Circuits and Signal Processing*, 13(1-2):123–151, May-Jun 1997.

[55] R. Sarpeshkar, W. Wattanapanitch, S. K. Arfin, B. I. Rapoport, S. Mandal, M. W. Baker, M. S. Fee, S. Musallam, and R. A. Andersen. Low-power circuits for brain-machine interfaces. *IEEE Trans. on Biomed. Circuits and Systems*, 2(3):173 –183, Sept 2008.

[56] H. Scherberger, M. R. Jarvis, and R. Andersen. Cortical local field potential encodes movement intentions in the posterior parietal cortex. *Neuron*, 46:347–354, April 2005.

[57] M. D. Scott, B. E. Boser, and K. Pister. An ultralow-energy ADC for smart dust. *IEEE J. of Solid-State Circuits*, 38(7):1123–1129, Jul 2003.

[58] M. D. Serruya, N. G. Hatsopoulos, L. Paninski, M. R. Fellows, and J. P. Donoghue. Instant neural control of a movement signal. *Nature*, 416(6877):141 –142, Mar. 14 2002.

[59] F. Shahrokhi, K. Abdelhalim, D. Serletis, P.L. Carlen, and R. Genov. The 128-channel fully differential digital integrated neural recording and stimulation interface. *Biomedical Circuits and Systems, IEEE Transactions on*, 4(3):149–161, Jun 2010.

[60] D. B. Shire, S. K. Kelly, J. Chen, P. Doyle, M. D. Gingerich, S. F. Cogan, W. A. Drohan, O. Mendoza, L. Theogarajan, J. L. Wyatt, and J. F. Rizzo. Development and Implantation of a Minimally Invasive Wireless Subretinal Neurostimulator. *IEEE Trans. on Biomedical Engineering*, 56(10):2502–2511, Oct. 2009.

[61] A. A. Sodagar, K. D. Wise, and K. Najafi. A fully integrated mixed-signal neural processor for implantable multichannel cortical recording. *IEEE Trans. on Biomed. Engineering*, 54(6, Part 1):1075–1088, Jun 2007.

[62] A. M. Sodagar, G. E. Perlin, Y. Ying, K. Najafi, and K.D. Wise. An implantable 64-channel wireless microsystem for single-unit neural recording. *IEEE J. of Solid-State Circuits*, 44(9):2591–2604, 2009.

[63] M. Steyaert, W. Sansen, and Z. Chang. A Micropower Low-Noise Monolithic Instrumentation Amplifier for Medical Purposes. *IEEE J. of Solid-State Circuits*, 22(6):1163–1168, Dec 1987.

[64] D. M. Taylor, S. Tillery, and A. B. Schwartz. Direct cortical control of 3D neuroprosthetic devices. *Science*, 296(5574):1829–1832, Jun. 7 2002.

[65] Y. Tsividis. *Operation and Modeling of the MOS Transistor.* McGraw-Hill, 2nd edition, 1998.

[66] C. Tung-Chien, L. Wentai, and C. Liang-Gee. Vlsi architecture of leading eigenvector generation for on-chip principal component analysis spike sorting system. In *IEEE Engineering in Medicine and Biology Society*, pages 3192 –3195, 2008.

[67] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz. Cortical control of a prosthetic arm for self-feeding. *Nature*, 453(7198):1098–1101, Jun. 19 2008.

[68] E.A. Vittoz and O. Neyroud. A low-voltage cmos bandgap reference. *IEEE J. of Solid-State Circuits*, 14(3):573–579, June 1979.

[69] W. Wattanapanitch, M. Fee, and R. Sarpeshkar. An Energy-Efficient Micropower Neural Recording Amplifier. *IEEE Trans. Biomed. Circuits and Systems*, 1(2):136–147, Jun. 2007.

[70] J. Wessberg, C. R. Stambaugh, J. D. Kralik, P. D. Beck, M. Laubach, J. K. Chapin, J. Kim, J. Biggs, M. A. Srinivasan, and M. Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408(6810):361–365, Nov. 16 2000.

[71] C. M. Zierhofer and E. S. Hochmair. Geometric approach for coupling enhancement of magnetically coupled coils. *IEEE Trans. on Biomedical Engineering*, 43(7):708 714, Jul 1996.

[72] X. Zou, X. Xu, L. Yao, and Y. Lian. A 1-V 450-nW Fully Integrated Programmable Biomedical Sensor Interface Chip. *IEEE J. of Solid-State Circuits*, 44(4):1067–1077, Apr 2009.

[73] Z.S. Zumsteg, C. Kemere, S. O'Driscoll, G. Santhanam, R.E. Ahmed, K.V. Shenoy, and T.H. Meng. Power feasibility of implantable digital spike sorting circuits for neural prosthetic systems. *IEEE Trans. on Neural Systems and Rehab. Engineering*, 13(3):272–279, Sept 2005.

[74] A. Zviagintsev, Y. Perelman, and R. Ginosar. Algorithms and architectures for low power spike detection and alignment. *J. of Neural Engineering*, 3(1):35–42, Mar 2006.