Title
A System-Level Analysis of a Wireless Low-Power Biosignal Recording Device

Author
Chandler, Rodney James

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

A System-Level Analysis of a Wireless Low-Power Biosignal Recording Device

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Rodney James Chandler

2012

Abstract of the Dissertation

A System-Level Analysis of a Wireless Low-Power Biosignal Recording Device

by

Rodney James Chandler

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2012

Professor Jack W. Judy, Chair

Development of brain-machine interfaces and treatment of neurological diseases can benefit from analysis of recorded data from implanted electrodes. Existing wireless neural recording systems are often bulky, dissipate too much heat to be implanted, or only have a small number of channels. Furthermore, advances in micro-machined electrodes provide the possibility of high-density recordings, but the companion electronics do not provide enough simultaneous channels with low enough power, wireless telemetry, or a small form-factor. A system level view of wireless recording-circuitry which could overcome these deficiencies is described in this work. The overall system comprises of an analog front end (AFE), digital signal processing (DSP), and transmitter (TX). Each block is analyzed, and system-level specifications are derived. Based on these specifications, each block can be optimized for low power and small area. The analog front-end uses open-loop amplifiers to support lower voltage operation than previously published work. A prototype amplifier was also fabricated to measure performance in a 65-nm CMOS process that is needed for low-power digital signal processing. The amplifier performance was comparable to other recently published amplifiers with 2.5 μV noise in 10 kHz bandwidth while dissipating 17.2 μV from a low 1 V supply. The use of programmable bias currents in the amplifier, to exploit the trade-off between noise and power, was proposed to set each individual amplifier's noise level (and power) to meet

requirements for accurate spike detection. Literature reviews of digital-signal processors and transmitters are used to construct approximate models of power versus performance. These models are then used to investigate the overall system power with different levels of digital processing. With a target application of neural spike recording, four modes (raw data, spike detection, feature extraction, and clustering) were analyzed. A system that uses feature extraction yields the lowest overall power, supports 400 channels with a practical wireless link, and consumes approximately 8 mW.

The dissertation of Rodney James Chandler is approved.

Hugh T. Blair

Dejan Marković

Sudhakar Pamarti

Jack W. Judy, Committee Chair

University of California, Los Angeles

2012

To Karen...

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Professor Jack Judy for advising me over the course of the Ph.D. program. His thorough criticism and word-smithing of my work has been beneficial, and has helped me to communicate my ideas more effectively.

I would also like to thank Professors Sudhakar Pamarti, Tad Blair, and Dejan Marković for serving on my committee, and for their valuable comments in refining this manuscript. In particular, I give many thanks to Professor Marković for virtually serving as my co-advisor while Professor Judy was on sabbatical, and supporting joint chip tapeout for this work.

I also thank my friends at UCLA. Shahin Farschi, Rocco Tam, Sarah Gibson, Victoria Wang, Mansour Rachid, Cathal McCarthy and David Murphy are just a few of the people that I've been fortunate to meet. I am grateful for the many technical discussions that take place in Professor Judy's group, particularly with June Isobe, Frank Zee, and Jere Harrison. I am indebted to Beverley Eyre for explaining many LaTeX techniques.

I greatly appreciate the continuous help and support from Kyle Jung, Deeona Columbia, Martha Contreras, Alice Brook and Ilhee Choi. They were always very helpful with numerous urgent requests.

Lastly, I thank my wife, Karen, for her endless love, support and proofreading. I could not have done this without her.

Vita

1998	B.E., Electrical and Electronic (Honours I), The University of Queensland
1999	Graduate Student Researcher, The University of Queensland
2000-2003	IC Design Engineer, Radiata/Cisco Systems, Sydney
2003-2010	Graduate Student Researcher, University of California at Los Angeles
2007	IC Design Engineer, Analog Devices, Wilmington, MA
2008	M.Sc., Electrical Engineering, University of California at Los Angeles
2010-2011	IC Design Engineer, MaxLinear, Carlsbad, CA
2012-	IC Design Engineer, Broadcom, San Diego, CA

Publications

R. Chandler, S. Gibson, V. Karkare, S. Farshchi, D. Markovic, and J. W. Judy, "A System-Level View of Optimizing High-Channel-Count Wireless Biosignal Telemetry," in Proc. Int. IEEE Engineering in Medicine and Biology Conf., Sept. 2009.

S. Gibson, **R. Chandler**, V. Karkare, D. Markovic, and J. W. Judy, "An Efficiency Comparison of Analog and Digital Spike Detection," in Proc. 4th Int. IEEE EMBS Conf. on Neural Engineering, May 2009.

D. Huber, **R. Chandler**, and A. A. Abidi, "A 10b 160MS/s 84mW 1V Subranging ADC in 90nm CMOS," in IEEE Solid-State Circuits Conference, 2007.

CHAPTER 1

Introduction

1.1 Background

Neuroscience research is required for several reasons. This work, in particular, has been influenced by the desire for (a) treatments for neurological disorders requiring measurement and monitoring of abnormal behavior of the brain, (b) developing models for the normal function of the brain, so that synthetic circuits can be devised to restore function to damaged regions, and (c) develop brain-computer interfaces (examples are implementing brain-controlled prosthetic limbs for amputees, more intuitive, natural, machine-control interfaces, and restoring motor control to stroke victims).

All the above examples rely on monitoring brain signals. These signals can be monitored in a variety of ways, from electroencephalography (EEG), which requires electrodes placed on the scalp, to micro-electrode implantation in the brain. Depending on the technique employed, the data gathered will range from the ensemble response of millions of neurons (EEG), down to the behavior of a handful of individual neurons per implanted electrodes. EEG-based monitoring is suitable for systems that can be trained to detect high-level behavior such as blinking or the intention to move in a certain direction. On the other hand, it is theorized that signals recorded from implanted electrodes can be used to obtain more precise information. One study of the combined network interaction from hundreds of neurons is led by Blair *et. al.*. The specific aims of Blair's work are to (a) measure how position and velocity are encoded in the rat brain, within specialized neuronal networks known as

grid- and place-cells, and (b) develop mathematical models for this behavior. Furthermore, long-term goals include applying their results, based on the mapping of physical quantities such as position and velocity, to more abstract concepts such as how one navigates through a lifetime of memories.

As many preclinical trials use small-animals such as mice and rats, the size and power of the system must be minimized. A small, low-power system facilitates a broad range of applications (from first-order preclinical small-animal trials to primate experiments). All applications require similar processing, but differ greatly in their power, size, and weight constraints. Therefore, it behooves us to develop the system with the tightest constraints in mind. A platform that is flexible enough to be used in many neuroscience applications, such as the one proposed in this dissertation, would encourage the sharing of collective experiences, and assist in developing best practices for neural recording, while also allowing neuroscientists to focus on the experiment itself.

Before describing the requirements of the proposed system, a few applications will be presented to set the stage.

1.2 Applications

Successfully developing the proposed technology will make it possible to routinely perform wireless high-channel-count and high-bandwidth neural recording in freely moving test subjects. Such a capability has important clinical (e.g., epileptic seizure mapping) and preclinical (e.g., brain-computer-interface development and recording within enriched environments) applications.

1.2.1 Clinical Application: Epileptic-Seizure Mapping

Epilepsy and its development in the intact brain are poorly understood. The gross features of the disease are known, but the nuances of its workings, its development, and its natural triggers remain largely obscure. In broad strokes, an epileptic seizure occurs when neurons in a small area of the brain begin to fire in synchrony. This firing overwhelms other necessary brain functions and results in a complete shutdown of the central nervous system and a loss of consciousness in extreme cases. Epilepsy affects millions in the United States alone — according to some estimates as much as 1 to 2% of the population [1]. Epileptic insults range in their effects from the facial twitching and momentary blankness of expression typical of petit mal seizures, to the emotional outbursts, hallucinations, and flashbacks of temporal-lobe seizures, to the loss of consciousness and violent convulsions seen in grand mal seizures. Causes of epilepsy include genetics, head trauma, and various brain disorders. Epilepsy is normally treated with medications designed to inhibit neuronal activity by increasing the effectiveness of the inhibitory neurotransmitters. There is also on-going research into using electrical stimulation to manage the disease.

Scientifically, one of the most important questions to answer relates to the process by which the brain develops the capacity for epileptic insults, before any overt signs of epilepsy are present. Clinically, the goal is to localize the source of drug-resistant seizure activity to enable surgical intervention. When such localization cannot be performed with non-invasive imaging technologies, direct brain electrophysiology may be used. Presently, patients are admitted to the epilepsy ward, surgery is performed to attach a grid of many electrodes to the cortex or to implant electrodes into the brain, and continuous recordings are performed in the clinic until enough seizure activity is recorded to enable accurate source localization. Unfortunately, a well-known fact is that seizure frequency during such clinical experiments is far less than normal. As a result, patients often must stay in the hospital for up to two weeks for enough data to be obtained [1]. Such a long hospital stay with many (~100) transcranial

and percutaneous wires is a great burden to the patient and increases their health risks. A hypothesis for the cause of reduced seizure activity observed in the clinical setting is the fact that the physical activity and behavior of patients is greatly reduced and altered due to being tethered to the bed. In addition, the wired nature of existing systems means that recordings are not obtained when the patient moves away from the bed (e.g., moving between rooms, etc.) and it is possible that the already infrequent seizure activity may be missed. A wireless system capable of recording from ~100 channels may provide substantial benefits to the patient. Although typically the system may only need to record signals with a sampling rate of <1000 Hz, a system capable of sampling at higher rates (~20 kHz) would enable the same system to simultaneously record from implanted micro-electrodes. Furthermore, the ability of the mobile wireless system to perform local signal processing could enable the early detection of seizure onset, which would also have clinical benefits (i.e., prepare the patient and care provider for the seizure so that the data obtained can be of the highest value).

1.2.2 Preclinical Application: Brain-Computer Interfaces

The development of brain-computer interfaces (BCI) is an active area of research based on multi-channel EEG [2] or single-unit electrophysiological recordings [3–9]. A goal of BCI research is to restore movement to those who have lost a limb, or link a computer interface to the brains of those suffering from cerebral palsy or Lou Gehrigs disease. By recording the electrical activity from the brain, a computer can be programmed to deduce the intention of movement and control a robotic device or prosthetic. Although the early work performed with rodent models used a fair number of electrodes (~16), there has been a drive towards using even larger numbers. For example, there has been success with BCI systems operating with ~100 channels non-invasively with humans, and invasively with a higher-order preclinical animal model (i.e., primate) [9]. Although a wired system is acceptable for such experiments that are typically performed acutely and in a configuration that is physically constraining, there are experimental motivations for a wireless neural-recording system. One motivation is

to study the long-term performance of invasive neural-electronic interfaces, which are known to degrade with time sometimes slowly (over months) and sometimes quickly (sub-second) [10]. Another motivation is to perform synchronous and asynchronous BCI experiments in mobile test subjects, initially in animals [5] [6] but then eventually in humans. With non-invasive BCI recording systems, a low sampling rate (<1000 Hz) is sufficient to capture EEG activity and the level of signal processing varies.

With invasive BCI recording systems, a high sampling rate (~20 kHz) is needed to capture spike activity, but there is also interest in recording local field potentials with the same system. Although some researchers may configure systems that record spike activity to simply stream the full waveform to the receiver, others may prefer the mobile system to perform substantial signal processing as well (e.g., event/spike detection and spike sorting). The ability of our proposed system to be reconfigured through software to employ a wide range of sampling rates, resolutions, and user-defined signal-processing algorithms, will make it extremely useful for the diverse BCI community, and not just for those using smaller animal models.

1.2.3 Preclinical Application: Recording of Brain Activity in Enriched Environments

The influence of an enriched environment on brain development, plasticity, and recovery is known to be critical but is poorly understood. The environment in which the animal lives and is studied, from the moment it is born, plays a critical role in initial and ongoing brain development.

Several paradigms have been used to demonstrate the effects of environment on lower-order animal models (e.g., rodents). The best characterizations of these differences have been reported in maternal behaviors [11], light-dark rearing [12], exercise [13] [14], diet [13–16], and environmental complexity [17–20] all of which induce lasting changes in brain anatomy

and/or function. One of the oldest models of experience-dependent plasticity is housing in an enriched environment. Bennett et al. were the first to report that rearing rats in complex housing conditions resulted in changes in brain weight and cortical thickness [19], as well as in behavior. In general, enriched-environment rearing provides frequent novel stimuli to test subjects in the form of toys, exercise devices, spatial arrangements, handling, social interactions, and sometimes even smells and sounds. When rats reared in an enriched environment are compared to those reared in standard laboratory conditions, or to those reared in "impoverished" conditions, the enriched-environment animals consistently demonstrate changes in neuroanatomy (e.g., increases in brain weight [19][21], cortical thickness [22], dendritic arbors [23], glial number [20], brain capillaries [24][25], and even hippocampal neurogenesis [26]), changes in behavior (e.g., superior cognitive performance on tasks of spatial learning, including Morris water maze [26, 27] and the radial arm maze [28]), and changes in pathological processes with high clinical relevance, including the recovery from a stroke or head trauma, to epilepsy, to the effects of pre-natal alcohol exposure.

It has long been contended that standard laboratory-animal-housing conditions are actually impoverished, when compared to 'wild type' surroundings [29]. In fact, it may be argued that comparison of experimental effects in standard-housed animals provides only part of the picture, particularly when attempting to extrapolate such results to the human condition. Important conclusions derived from such research demonstrate that differences in brain signaling as a function of the animal environment are sufficient to call into question conclusions made from research that is not done in an enriched environment [30].

An enriched environment is one that consists not only of a greater number of objects (e.g., wheels, tunnels, rocks, and toys), but also of an enriched social setting with additional animals. The optimal environment from a neuroscientific point of view is not the one with the most stimuli or animals, but the one that most resembles the natural environment of an animal in the wild. Overcrowding degrades neural development in much the same manner as does a lack of social interaction. Neurons in brain slices prepared from animals exposed to

an enriched environment show a dramatic increase in cell proliferation, dendritic branching, and expression of the genes that lead to receptor formation [30]. In most societies, it is fair to consider the normal human condition to be one of considerable 'enrichment', with complex social interactions, educational programs, language, art, and recreational opportunities. In the laboratory, it is likely that subtle but important impairment of "higher-level" functions will be missed by experimental paradigms that reduce animal behavior to the lowest common denominator.

A wireless neural-recording system that could be used to perform preclinical experiments with lower-order animals models as small as a rat, could be used to ascertain the importance of an enriched environment on conditions of clinical relevance (e.g., recovery from a stroke or head trauma, epilepsy, the effects of pre-natal alcohol exposure). Such a wireless neural-recording system may not initially need to record from a high channel count, but the availability of micromachined microprobe arrays makes it increasingly common to perform high-channel count experiments even in rats. Thus, the expandable performance of our proposed telemetry system is likely to be highly valuable. In addition, the frequency range of neural signals of interest in such experiments varies from <100 Hz (e.g., theta waves) to ~6 kHz (e.g., single units). The ability of the same miniature neural-recording system to be reconfigured in terms of channel count, resolution, sampling rate, and extent of signal processing would enable a greater range of experiments to be performed in an enriched environment.

1.3 System Description

Advances in miniaturization of microelectrode arrays have made it possible to chronically implant hundreds (or even thousands) of extracellular electrode tips into the nervous tissue of animal or human subjects in order to monitor the activity of dozens or hundreds of individual neurons in real time. In principle, it should be possible to gather data from these minia-

turized electrode arrays while unrestrained subjects move freely about their environment, thus capturing neural activity during normal behaviors. But in practice, this potential for high-density neural monitoring in unrestrained subjects has yet to be fully realized, because with hundreds or thousands of electrode channels implanted in a single subject, it becomes necessary to transmit very large volumes of real-time data from the brain to a machine that can process and store the data. A reliable high-bandwidth data-transmission bus is needed for relaying all of this data from brain to machine. In many cases, exquisitely miniaturized electrodes must still be connected to recording and monitoring equipment by large, bulky cables that restrict the ability of subjects to move and behave freely. Unavailability of interface circuitry that can multiplex hundreds of channels of data in a small form with low power have prevented miniaturized electrode arrays from being used to their full potential in applications such as brain-machine-interface devices for controlling robotic prosthetics in movement-disabled patients, and in laboratory research investigating how populations of neurons encode information and regulate behavior in freely behaving animals.

Many existing systems use a fully parallel data bus: a multiconductor cable in which each channel of data is relayed from brain to machine via a separate wire (Figure 1.1(left)). However, there are several major drawbacks to this solution. First, with hundreds or thousands of channels, the multiconductor cable can become very large and bulky, especially for small experimental animals like mice and rats. Second, the connectors that join these cables to the implanted microelectrode array are a very vulnerable point in the circuit, often consisting of hundreds of very small and fragile connector pins that can be easily broken during the process of connecting and disconnecting the cable to the subject (especially animal subjects, such as a squirming rat). If even a single connector pin is broken, a repair job costing hundreds of dollars and several man hours becomes necessary. Third, the connector cable tethers the subject to the data acquisition equipment, thereby restricting the subject's ability to move around and behave freely without twisting or breaking the cable. This problem can be partially ameliorated by using a rotational commutator, but such commutators

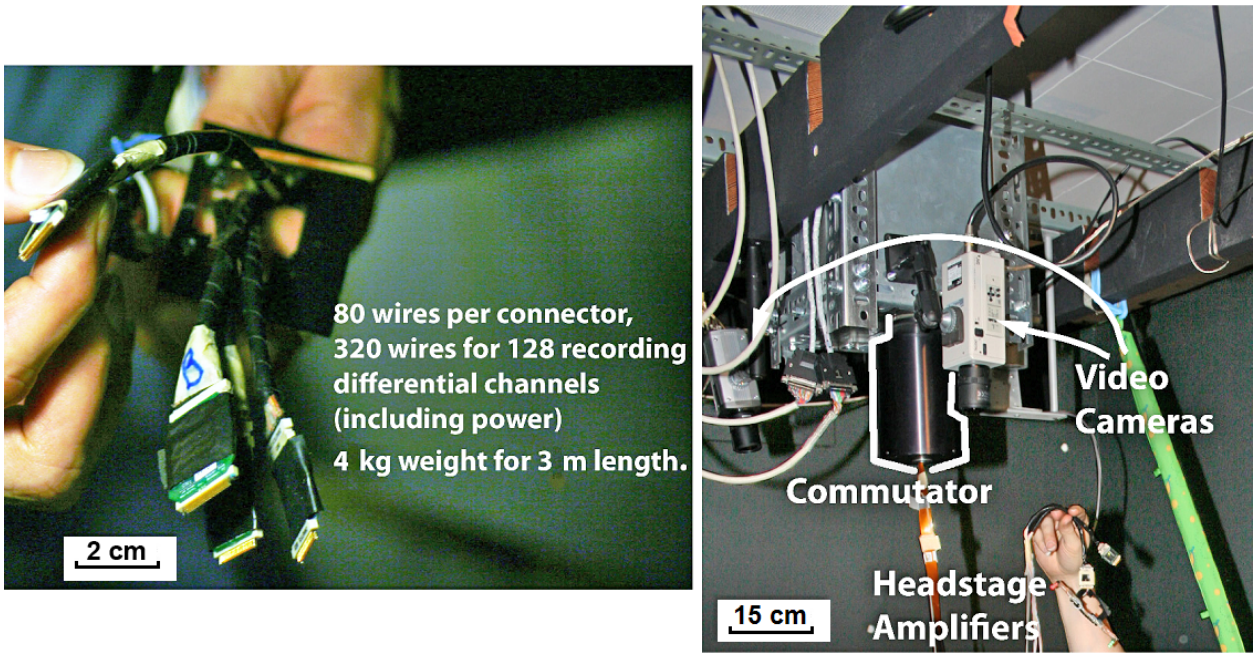

Figure 1.1: (left) Connector and cable detail. (right) Ceiling mounted commutator and video cameras, attached to headstage amplifiers via 3 m cable.

become very large and expensive for cables consisting of hundreds of conductors (Figure 1.1 (right)).

To solve these problems, we will develop a high-bandwidth, miniaturized wireless data transmission system for relaying large volumes of electrophysiological data from implanted electrode arrays to a data acquisition system. In collaboration with experienced users of existing technology, we have identified the major technological hurdles that must be overcome before such a wireless neural recording system can become truly practical and useful. Our system will solve these problems by incorporating a miniaturized, wearable data-acquisition computer system that can be remotely programmed by the experimenter at any time. We are confident that this system will make it possible, for the first time, to achieve ultra-high-density electrophysiological monitoring (hundreds or thousands of channels) of single-unit electrodes in fully unrestrained human and animal subjects.

The hardware to perform the neuroscience research described above has the following minimum requirements: (a) high-fidelity, simultaneous recording from hundreds of electrodes, (b) relaying the recorded signals to a digital-signal processor to perform experiment-dependent operations, such as detection, alignment, feature extraction, and classification, and (c) transferring the neural information to an archive server, or generating a electrical feedback signal.

To make the system implementable in practice, the following requirements are also introduced:

The system must be low power, to avoid tissue damage from heating, and to allow a low-weight battery for a given operational lifetime. Specifically, the total power dissipation must be less than 80 mW/cm^2. For a 100 channel, 1 cm^2 integrated circuit (IC), this upper-power limit corresponds to a power dissipation of less than 800 μW per channel. For a system weight of 10 g, assuming 50% is due to the battery with a energy density of 120 W·hr/kg (typical for lithium ion batteries), and a target experiment duration of 1 hour, the maximum power per channel is 120 W·hr/kg $\times$ 5 g / 100 channels, or 6 μW. In this example, we are limited by the power source, rather than limits due to potential tissue damage. Providing power over an inductive link is unsuitable because of its short range (~10 cm).

While it is arguable whether the system *must* be wireless, the key benefit of a wireless approach is the elimination of the wiring harness and the connectors. This harness and connectors are problematic because the system must be physically small, the wiring harness consists of low-gauge wire, and miniature connectors are fragile. The main problems that arise are (a) damaging the connector after numerous reconnections and (b) the weight and limited movement due to the wiring itself. A small, squirming rat only adds to the likelihood of a damaged wired link. As expensive, custom circuitry and hand-made housings are presently used in the experiment, time-consuming repairs are undertaken instead of replacement. As surgery to implant electrodes is also time-consuming and expensive, it is imperative to minimize damaging the hardware attached to the implants. The heavy cable

must also be towed by the rat, which impedes the extent of their mobility, and eliminates the possibility for experiments using *enriched* environments including tunnels and complex mazes. A rotational commutator improves mobility, but ideally, the rat would be untethered.

1.4 Proposed Solution

This work aims to deliver a device that is small, yet capable of wirelessly transmitting as much information as possible with a given battery. Some investigators seek real-time, raw neural signals captured with high-fidelity (i.e., at least 8-bit resolution and 30 kS/s sampling rates), while others are only interested in the spike waveforms and/or events. The ability to transmit over distances up to 10 m, operate for up to 24 hrs, and yet weigh less than 10 g, would facilitate unencumbered movement for the subject, longer experimental sessions, and the ability to use the system with small animals, respectively. A system that can achieve these goals has the potential of broad appeal to the neuro-scientific community.

1.5 Organization

The remainder of the thesis is organized as follows. Chapter 2 summarizes previously published systems. In order to facilitate system level analysis, each building block is analyzed in Chapter 3. Next with the block estimates, the system level performance can be calculated for different configurations, and the optimal configuration determined. One of the key blocks is the analog front-end amplifiers, and the implementation of these amplifiers is described in Chapter 5. Finally, conclusions and future work are summarized in Chapter 6.

1.6 Acknowledgments

Work presented in Chapters 3 and 4 includes contributions from Sarah Gibson and Vaibhav Karkare for the algorithm and digital system design, which have previously been published as conference papers [31, 32]. We also thank the National Science Foundation for financial contributions (Grant No. DBI-0456125 and EECS-0824275).

CHAPTER 2

Literature Review

2.1 Introduction

Each wireless biosignal recording system is composed of several key parts: electrodes, amplifiers, detection and signal processing, and a transmitter. A complete system is configured by using a certain approach for each of the blocks, and partitioning the system into a front-end and a back-end. The front-end performs basic signal processing, and the back-end performs additional processing and archival of data. The resulting system partition leads to a certain level of performance in terms of power, noise, range, area, and datarate. Signal fidelity (e.g., distortion of the amplifiers and spike detection accuracy) is also important, but not always reported, and so it is not included here. Integrated circuits are low in weight, which indicates that the total system weight is mostly due to the battery, antennae and printed circuit boards. The required battery weight is related to power dissipation and experiment duration, so this constraint is implicitly captured. In this Chapter, we will review some existing systems in terms of these performance metrics. In Chapter 3, we will examine these performance trade-offs, e.g. power versus noise, in more detail.

Existing wireless neural-recording systems range from fully integrated analog transmitters [33],[34], to analog transmitters with threshold-based spike detection [35], [36], [37] to digital application-specific integrated circuits (ASICs) [38], to microcontroller-based embedded systems [39], to commercial-off-the-shelf (COTS) PC-based systems [40]. Fully integrated transmitters and ASICs benefit from being very small (several mm^2) and low power (sev-

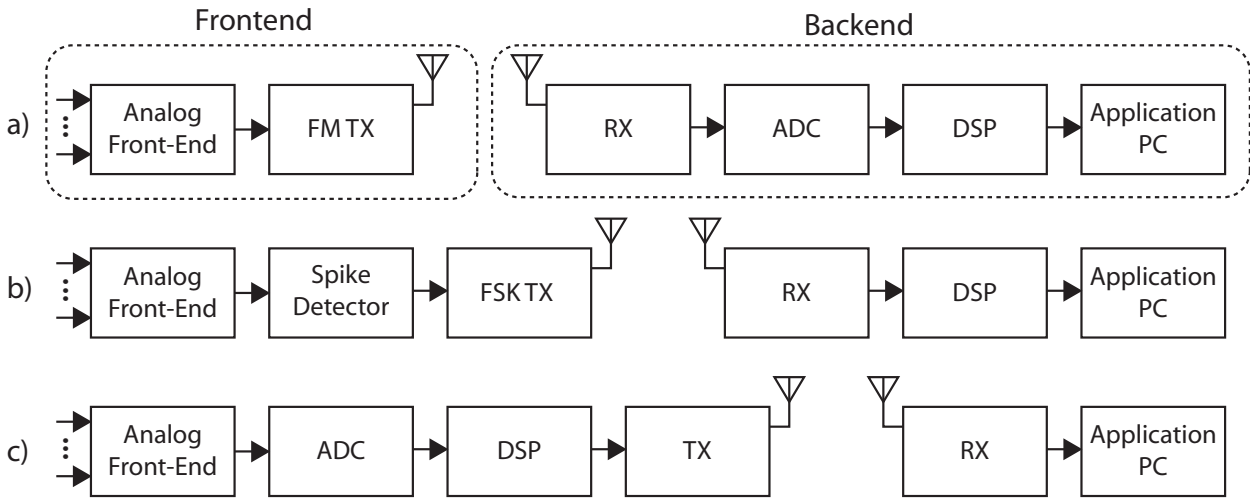

Figure 2.1: Different system configurations for wireless neural recording: (a) analog recording (b) spike detection (c) digital signal processing.

eral mWs), thus enabling them to be implanted with the electrode and inductively powered (when an external power source can be worn).

In the interest of increasing channel count and sampling rate while maintaining reasonable battery life, some groups have demonstrated solutions with some on-board signal-processing capability, such as thresholding as demonstrated in [35] and [36]. Unfortunately, these threshold-based systems typically cannot differentiate spikes from artifacts, and require circuit redesign to modify the spike-detection algorithms. The limited adoption of existing wireless neural recoding systems by the neuroscientific community may be an indicator that users could benefit from a greater degree of flexibility in terms of methods for spike detection. Chae has published a system in [41] that also includes feature extraction and clustering, and a high datarate transmitter. Such a system, with the capability to provide spike event data along with many channels of raw data, is a step closer to providing the needed flexibility.

2.2 Classification of Recording Systems

Figure 2.1 shows systems that use (a) analog recording, (b) spike detection, and (c) using a digital-signal processor (DSP) for feature extraction and/or clustering. Each system is comprised of two parts. The first part shown is the frontend, which includes the blocks from the analog frontend (AFE) to the transmitter (TX). The frontend takes inputs from several electrodes and applies the appropriate gain and frequency-band selection, and combines the multiple electrode signals into a single stream. The combined stream is transmitted wirelessly to the backend receiver. As the frontend is intended to be battery powered and/or implanted, it must be low power and have a small size, and any increase in frontend complexity must be balanced against these constraints. One of the key goals of this dissertation is to determine the optimal frontend complexity and explore the performance tradeoffs for different applications. The backend, which includes the blocks from the receiver (RX) to the application PC, does not have these constraints, so the design of the backend is relatively trivial compared to the frontend. The backend receiver can use off-the-shelf components, which suggests that the frontend transmitter should be able to interface standard components.

For the analog recording case (Fig. 2.1a), time-division multiplexing is used to interleave multiple electrode channels. Analog recording system examples have been published by Mohseni [33] and TBSI [42]. The basic architecture is shown in Figure 2.1a. The advantage of this type of system is simplicity. The amplifier increases the signal level from a few millivolts to approximately 1 V to drive an FM modulator. Because the output is an analog modulation, any noise and interference that is incident on the receiver can reduce the signal quality. Since there is no signal processing, it is not possible to detect spikes or other events in the frontend – signal processing occurs after the backend receiver. If a feedback signal is required (e.g. to control a prosthetic limb), it must to calculated in the backend, and transmitted back to the frontend. If the system included signal processing in the front-end,

it is feasible that the feedback signal could be generated locally. For this reason, analog recording has limited applications for applications that also include stimulation.

A spike detection system (Fig. 2.1b) is comprised of the following blocks. The amplifier and bandpass filter processes the signal to remove the energy outside the wanted band and increases the signal amplitude to be larger than the noise floor of subsequent blocks. A simple analog comparator then compares the instantaneous amplitude to a predetermined threshold and generates a pulse if the amplitude is higher. The output is now a 1-bit sequence, and this digital data can drive an FSK modulator instead of an FM modulator. Because the output is low resolution (1 bit), the transmitter complexity is low but the waveform has been reduced to a 1-bit representation. The data from multiple spike detectors are interleaved and synchronization markers are added before being wirelessly transmitted to the backend. Since the waveform is reduced to a single bit, no analog-to-digital converter (ADC) is required for the spike detection system. Examples have been published by Sodagar [43] and Harrison [35], which use a simple form of spike detection. Their spike detection operates by comparing instantaneous amplitude with a predetermined threshold. The threshold is set by measuring the average amplitude over an extended time period, which is a measure of the system noise floor, and setting the threshold to be larger than this noise floor. The threshold is typically programmable, and can be set by circuitry operating in the background [44], or by user intervention. How accurately can an actual spike event can be detected in the presence of noise? Analysis of this topic has been performed by Gibson [45], and it was shown that simple thresholding has inferior spike detection performance in high background noise environments. In turn, this leads to higher power dissipation of the frontend analog amplifiers to keep their own noise low. For this discussion, the key disadvantage to note is that spike detection should be as accurate as possible, and not use an inferior algorithm, in order to encourage widespread adoption of the system in neuro-scientific community. Incorporating a range of spike detection algorithms is one solution, as is including the ability to keep more raw data. Both Harrison and Sodagar include the capability to transmit the complete digitized

waveform of a user-selectable channel (with 8- or 12-bit resolution), in addition to 1-bit spike detection data. Adding the high-resolution waveform data to the thresholded outputs increases the data-rate that the transmitter must support at the cost of performance, but is a good compromise between only having 1-bit data for a high number of channels, or a small number of channels with high resolution. Furthermore, the single high-resolution channel can be used during initial setup and training to observe which electrodes have useful data and to set the threshold of the comparators. Spike detection relies on the fact that it is sufficient to record the time of a spike event, and discard the spike waveform details as unimportant. However, false alarms or missed detections will occur; this behavior needs to be acceptable to the experimenter, and the complexity (i.e. power and area) of the detector must be balanced with spike detection accuracy. If one compares analog recording and spike detection, spike detection allows a reduction in datarate required of the wireless transmitter or a higher number of channels to be transmitted. For these reasons, spike detection is a useful technique in reducing the power dissipation of the system. On the other hand, analog recording keeps the complete spike waveform and the background noise (which can be required for certain experiments) but it is more difficult to create a system with a high channel-count.

Finally, Fig. 2.1c shows a system that includes some form of digital-signal processing to process the raw data to obtain a reduction in the amount of data sent to the transmitter. This reduction in datarate is required for high-channel-count systems to keep the transmitter feasible within a limited power budget. Spike detection is performed first and then additional processing is performed. Feature extraction measures the key features of a spike waveform such as peak, slope, and duration. Clustering is an additional step that, based on certain waveform metrics, maps a spike waveform to a certain event or a single neuron. Feature extraction and clustering are computationally expensive, and need high resolution digitization to be accurate. Typically 10 to 14 bit resolution is used depending on the application. Feature extraction and clustering can be considered as a form of application-specific lossy

compression and they reduce the data-rate required of the transmitter; other compression general purpose techniques could be used but would be less efficient. For a high-channel-count wireless system, if the data is not processed on chip, the outgoing data-rate can be high. In the system published by Chae [41], the data-rate per channel is 180 kbps, or 18 Mbps for 100 channels. Generally, transmitters that have power dissipation limited to a few tens of milliwatts can achieve up to 10 Mbps in indoor environment. The ultra-wideband (UWB) transmitter implemented in [41] claims a peak data-rate of 90 Mbps, which is enough for 500 channels of raw data, over a short transmit range, while consuming 6 mW from the power supply.

There are three interesting claims in Chae's publication: (a) with optimization and low-power design techniques it is possible to produce a system with a high-channel count without resorting to compression and keeping the total system power around 6 mW, (b) it is possible to design a high data-rate transmitter that has a low power of 1.6 mW and a capacity of 90 Mbps to support a high-channel count, and (c) feature extraction in the frontend can also have a low power of 1.5 mW for 100 channels. Overall, it shows that a low-power neural recording system is feasible based on (a) and (c). Regarding (b) however, the implementation of the wireless link and transmitter may be difficult. My own opinion is that the UWB transmitter performance may be sensitive in realistic settings. In particular, our goal is to have a range of approximately 10 m to allow untethered recording of freely-moving subjects. It is unclear if the UWB link is robust enough to supply high data-rate that is claimed at longer range and with other potential wireless interference nearby. Since the Federal Communication Commission dictates that UWB transmit power levels should be low, there may be issues with achievable link quality. Furthermore, widespread adoption of UWB technology in general consumer markets has been limited, with other non-UWB radio products satisfying the need for high datarate wireless communications. One possible reason for the lack of UWB adoption is that it may be difficult achieve high data-rates in real-world situations. Furthermore, obtaining off-the-shelf components for a UWB receiver

Table 2.1: Summary of existing wireless neural recording systems.

Author	TBSI	Mohseni	Sodagar	Harrison	Chae	Walker
Year	2007	2005	2007	2007	2008	2011
Ref.	[42]	[33]	[43]	[35]	[41]	[46]
Type	Analog	Analog	Spike/DSP	Spike/DSP	DSP	Analog
Channels	62	4	64	100	128	96
Power (mW)	30.8	2.2	14.4	13.8	6.0	6.5
Area (mm^2)	25	4.8	19	28	63	25
Noise (μW)	10	7.1	8	5.1	4.9	2.2
Bandwidth (kHz)	3	10	10	5	20	10
Datarate (Mbps)	4	4	0.50	0.33	90	30
Transmitter	FM	FM	FSK	FSK	UWB	N/A
Range (m)	3	0.5	0.02	0.13	1	N/A

may be more expensive than non-UWB options. Based on these concerns, I feel that it may still be necessary to use some form of data compression to keep the wireless link more robust and power dissipation reasonable.

Although the system does not provide wireless capability, Walker [46] has recently demonstrated a 96-channel system with a power dissipation of 6.5 mW while providing 10-bit raw data for all channels over wired serial connection. The design achieved low-power and noise by optimized analog amplifiers, and its techniques are applicable to all neural signal acquisition systems.

A summary of published solutions is shown in Tab. 2.1. In the next section, the performance of the each of published systems will be summarized, and compared to our goals for a low-power wireless neural recording system.

2.3 Summary

Key examples of wireless telemetry systems have been described in this Chapter. For this summary, we will examine several performance metrics: power, area, noise, and range. For power and area, we will calculate the approximate per-channel value by dividing by the number of channels. This is not the most accurate model as certain shared blocks such as reference voltage generators and local oscillators may not need to be scaled per channel. Noise is a measure of the minimum signal that can be detected, and is primarily set by the analog front end, i.e. power and area would increase if the noise is lowered. Noise often scales inversely with the square of power and area. We will consider noise in a qualitative manner in this section, and defer detailed analysis for Chapter 3. Range is also very application specific, as some systems only require a very short range if an intermediate relay stage is added between the frontend and backend. Other systems, on the other hand, were designed from the outset to be capable of longer range. Nonetheless, we can still obtain a qualitative assessment for what has been achieved to date. Our discussion focuses on scaling these systems to a large number of channels, hence we need to estimate the performance for a single channel. A summary of performance is given in the following list:

- Power: The analog recording systems in Table 2.1 use approximately 0.5 to 1 mW of power per channel. Mohseni uses approximately 140 μW per channel, plus 1.5 mW for the frequency modulator. Sodagar uses 225 μW per channel for the spike detection system. The FSK modulator in Harrison's system uses 6.9 mW, plus 69 μW per channel. The lowest power system is by Chae and dissipates 1.6 mW in the transmitter and 34 μW per channel. If we consider the per-channel power and examine Chae's system, we find that circuit techniques to disable blocks when their output is not sampled is used to reduce overall power; the effect switching in the signal path is not understood however and we feel that further analysis is required. Secondly, the UWB transmitter is much lower power per bit transmitted. Conservative estimates

would place the transmitter power between these extremes, and we expect around 5 mW for our transmitter. Our expected per-channel power, i.e. the amplifier and signal processing, is around 50 μW. Even in systems that use extensive digital signal processing, most of the power is used by the analog frontend and transmitter. For this reason, we will focus on reducing the power of these blocks and it may be advantageous to use additional DSP to reduce the power required of the amplifiers and transmitter.

- Area: The area-per-channel ranges from 1.2 mm^2 to 0.29 mm^2, and process technologies range from 1.5 μm to 90nm CMOS. The largest area-per-channel is from Mohseni with its unoptimized area as shown in the published prototype, is excluded from this area discussion. The smallest is by Harrison at 0.29 mm^2 despite it having the having the lowest noise, which implies that it is a well-optimized design. Both of the spike detection systems have small area, which is not surprising since they do not include the ability to digitize all channels simultaneously. The analog recording system by TBSI and the feature extraction system by Chae have areas of 0.40 mm^2 and 0.49 mm^2 respectively. Since the digital-circuitry area of all these systems is not dominant, process scaling (i.e. 65-nm CMOS) is of limited benefit. In fact the smallest area-per-channel system is implemented in 350 nm CMOS by Harrison. It is difficult to identify a clear trend from this data, and we summarize the following: the expected area is between 0.3 and 0.49 mm^2 per channel.

- Noise: The amplifier performance sets the limit on achievable noise, and can be traded off with power and area. The systems in Table 2.1 show 5 to 10 μV of noise. We will dissect this specification thoroughly in Section 3.3.

- Range: The range achievable is a function of power transmitted from the antenna and the data-rate, and will be examined in Section 3.7. The transmitter power dissipation has two main components, from the oscillator and upconverter/power amplifier circuits. Increasing range is primarily a function of the power amplifier, while the oscillator

power is set by required phase noise of the modulation scheme used.

One final point is that these systems have often been designed with a fairly specific application in mind. This makes a direct comparison difficult. One outcome of this dissertation will be to gain a better understanding of the implications of circuit imperfections such as analog circuit noise and distortion, ADC resolution, and feature extraction algorithms, on the overall quality of the recorded data.

In this Chapter we have examined system published to date to determine stat-of-the-art performance and generate some rough estimates of performance for our system. Chapter 3 deals with the optimization of individual blocks where these estimates will be scrutinized, while Chapter 4 deals with the overall system optimization.

CHAPTER 3

Block Level Design

3.1 Introduction

In this chapter we review the power and area estimates for each of the major blocks in the wireless biosignal telemetry system. Each of the blocks (neurons and electrodes, amplifier, analog-to-digital converter, digital-signal processors, and transmitter) shown in Figure 3.1 will be discussed in the following sections. The key concerns are also listed for each block.

3.2 Neurons and Electrodes

This section discusses the electrical characteristics of neurons and the electrodes used to sense their activity.

There are several classes of biological signals that can be measured, with different fre-

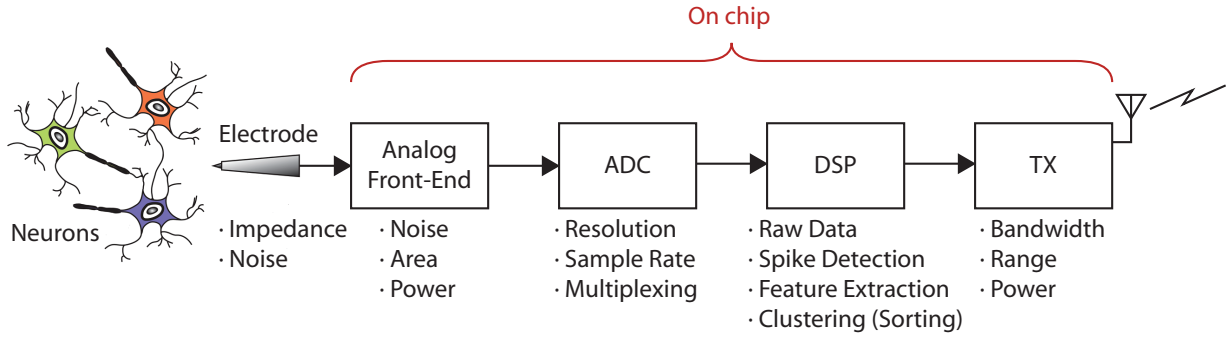

Figure 3.1: Block diagram of biological signal recording system.

quency content and amplitudes. Each type of signal use specialized electrodes, yet the signal level is still too small for signal processing. Amplifiers provide two main benefits: (a) a low level signal can be corrupted by additional noise in the signal processing blocks if gain is not applied and (b) the amplitude must also be increased to be compatible with the ADC full scale.

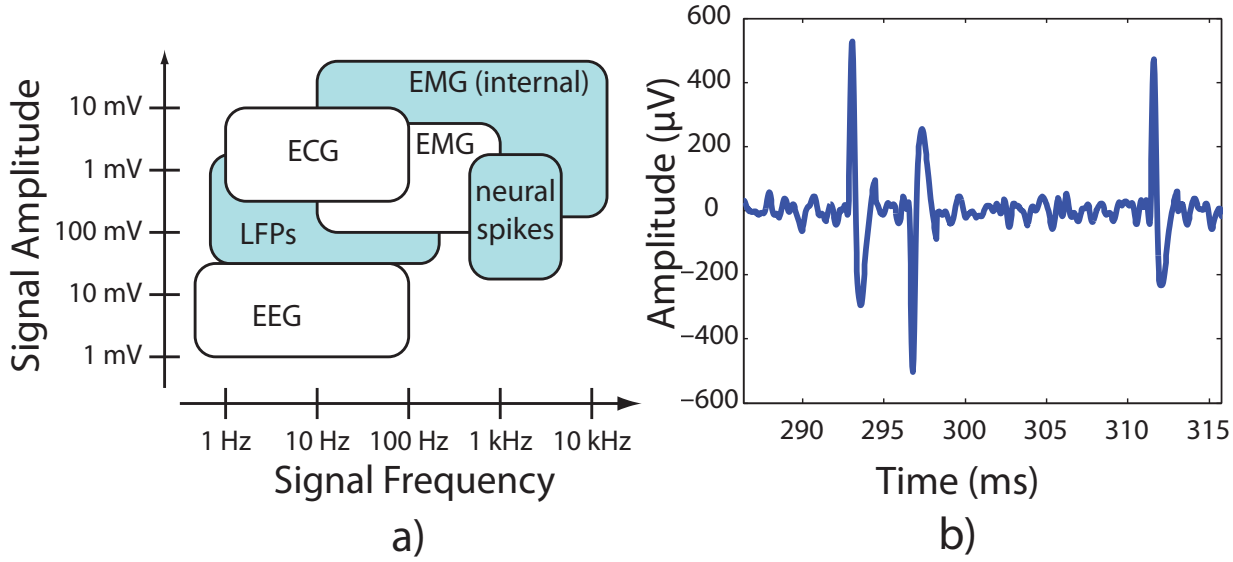

Figure 3.2: (a) Characteristics of different biosignals, (b) A recording of a neural spike.

Figure 3.2a shows commonly measured biological signals as described by Harrison in [47]. These signals are

- EEG: Electroencephalography is the recording of the ensemble behavior of tens of thousands of neurons within a brain, typically on the scalp. EEG is a large-scale observation of brain activity, and has poor spatial resolution.
- LFP: Local-field potentials are recorded from implanted electrodes, and sense the activity of neurons up to 3 mm from the electrode. A 300 Hz filter is used to remove higher frequency components.

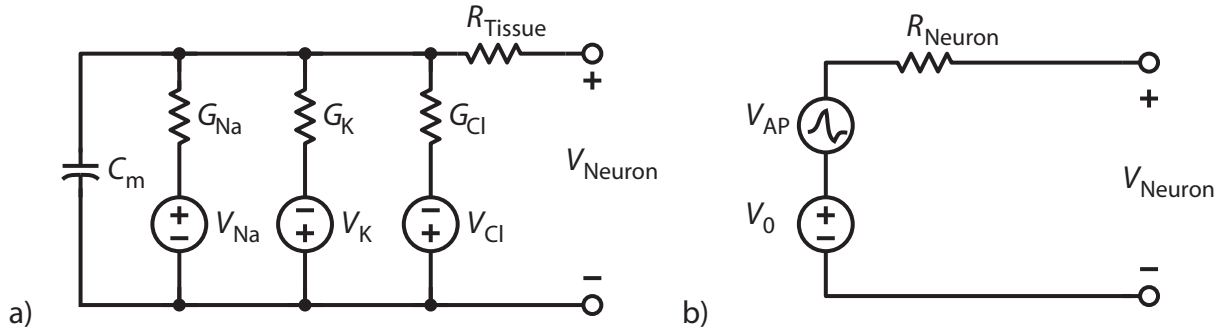

Figure 3.3: An Electrical Model of a Neuron.

- ECG/EMG: Electrocardiograph and Electromyography record the electrical activity of cardiac and skeletal muscles respectively.

- Neural spikes are recorded from an implanted electrode, capture the activity of a small number of neurons. Specifically, neural spikes refer to extracellular action potentials. Recording neural spikes is the main goal of this research. An example of a neural spike, after bandpass filtering to remove unwanted signals such as LFP, is shown in Fig. 3.2b. Two (possibly three) neurons can be identified by their spike shapes. The spike recording data was taken from Rutishauser *et. al.* [48].

From an electrical perspective, a neuron is comprised of a number of voltage-controlled ion channels, such as sodium, potassium, and chlorine. Each ion has an associated electrical potential, and conductance from inside the cell to outside. From a circuit perspective, the neuron is simply modeled by a voltage source with a dc offset V_{O} and an action-potential waveform V_{AP}. There is also a resistance associated with the neuron from the average conductance during a spike event, and a spreading resistance from the neuron to the electrode through the intervening ionic solution.

A simple electrical model of neuron is shown in (Fig. 3.3). The different ions channels are shown in (a). The ion channel conductances are voltage dependent, and the reaction time of these channels determines the shape of the action potential (also known as a spike).

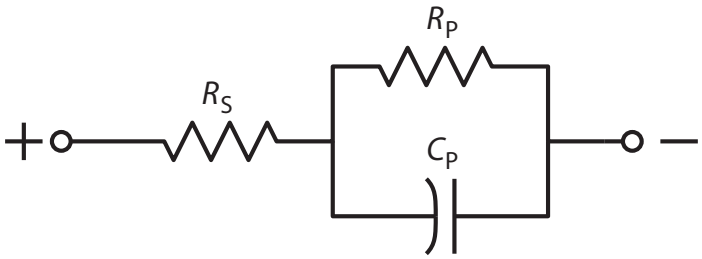

Figure 3.4: Electrode Model

A lumped model which reduces the network in (a) to an offset voltage, a transient voltage source, and a source resistance is shown in (b). It is assumed that R_{Neuron} is small compared the electrode resistances. For circuit modeling, (b) is adequate.

While the electrode itself is ideally little more than a conductor in most cases, the interaction between the electrode and electrolytic solution surrounding it requires modeling of the electrochemical interactions. Firstly, the electrode has a finite surface area, which introduces a resistive path R_{S} from the electrolyte to the electrode. Secondly, the charged ions at the interface form a voltage gradient, and this charge-storage can be modeled as a capacitor C_{P}, with a dc voltage across the capacitor representing the electrochemical interaction of the ions at the interface. The voltage across the interface can be up to 1-2 V [49],[50]. Proper selection of the ground electrode for the circuit can help cancel this offset voltage, and keep the inputs within the input-common-mode range of the amplifier.

Differential sensing also reduces the offset to the dc mismatch between the reference and sensing electrode. A resistive leakage path (R_{P}) also exists in parallel with the capacitor, but is typically negligible when considering the frequency response of the probe. However, it does show the dc offset of the electrode/electrolyte interface is coupled to the amplifier input. It should be noted however that the electrode interface is actually a complex non-linear system, and the model shown here is a simplification to enable circuit design. More

details can be found in [51].

3.3 Analog Front-End Amplifier

3.3.1 Differential Amplifier Configuration

In general terms, the neural amplifier must

- reject large dc offset voltages that occur at the tissue-electrode interface,
- provide enough frequency selectivity of the small neural signal of interest
- amplify the small signal to levels used by subsequent signal processors
- reject unwanted signals such as ac mains and other common-mode disturbances
- provide selectable reference potentials
- be able to drive the following stages (typically an ADC)
- have low input-referred noise

The dc offset voltage arises due to two main effects. The electrochemical half-cell that occurs around the conductive electrode in the ionic inter-cellular fluid gives rise to a voltage between approximately ± 1.5 V depending on the electrode material [50]. Since we detect the differential voltage between the desired electrode and a reference electrode, which helps ensure that the signal received by the amplifier is within its common-mode range, the amplifier only sees a fraction of the half-cell potential. The reference electrode must be connected to an appropriate potential. The potential outside a nerve cell also depends on the concentrations of the ions inside and outside of the membrane, and this resting potential is around -60 mV [52]. The combined effect of these dc offsets results in a few hundred millivolts of

dc offset at the amplifier input which must be removed before amplification, otherwise the linear range of the amplifier will be exceeded, and distortion and/or saturation will occur.

As seen in Fig. 3.2a, different signals occupy different frequency bands. To extract spike information efficiently, other signals such as LFP must be removed to minimize the required dynamic range of the amplifier. Spike frequency content is approximately within 1 kHz to 10 kHz. A filter profile with a high-pass corner frequency of 100 Hz and a low-pass corner of 10 kHz is used. As spike amplitudes for extra-cellular recording are typically between 50-500 μV, and ADC full-scales voltages are on the order of 0.5 V, we require a maximum voltage gain of 10,000 to match the ADC's full-scale. As the amplifier input is high impedance, it is susceptible to coupling of undesired signals such ac mains (50 or 60 Hz). For robust operation, amplifier operation should be largely unaffected by a 10 m$V_{\mathrm{p-p}}$ common-mode signal at 60 Hz. Specified another way, this signal should result in no more than 5% of the amplifier full-scale output swing, which gives a common-mode gain requirement of less than -52 dB. The amplifier should also provide a multiplexer to select different reference potentials. Such flexible configurations allow for optimizing common-mode noise rejection by choosing a reference electrode judiciously. This is because it is common to have access to a choice of reference electrodes, and it is not generally known in advance what combination of electrode and reference will give the cleanest signal. The amplifier must also be able to drive the ADC input impedance, which is typically a few picofarads for ADCs in the 8-12 bit resolution range.

Finally, the amplifier should have low input-referred noise. As noise and power are tightly linked, the noise specification will be carefully examined. The SNR is defined as

$$SNR = 20 \log_{10} \left(\frac{V_{\mathrm{p-p,spike}}}{6 \cdot v_{\mathrm{n,tot}}} \right) \tag{3.1}$$

where $V_{\mathrm{p-p,spike}}$ is the peak-to-peak voltage amplitude of the spike waveform, and $v_{\mathrm{n,tot}}$ is the input referred rms noise voltage of the system (of which the dominant sources are the electrode and amplifier stages).

Based on system level studies [53], an SNR of greater than 5 dB is required (Fig. 3.5), which in turn, specifies that for a 50 μV spike the total noise of the system should be less than 4.7 μV. Electrode resistance of 45 kΩ (per electrode in the signal/reference pair) gives 3.9 μV thermal noise integrated over a 10 kHz bandwidth, which then leads to an amplifier noise requirement of less than 2.5 μV. The noise contributions for 5 dB SNR for a low and high amplitude spike are shown in Table 3.1. The table also shows that a high amplitude spike can tolerate a much higher amplifier noise $v_{\mathrm{n,in}}$ for the same SNR. The associated power-noise trade-off will be explored in Section 3.3.7, where we consider a technique to minimize power dissipation.

Table 3.1: Noise Contributions for 5 dB SNR

	Low Amp.	High Amp.	Unit
BW	10000	10000	Hz
SNR	5	5	dB
$V_{\mathrm{p-p,spike}}$	50	500	μV
$v_{\mathrm{n,tot}}$	4.69	46.86	μV
$v_{\mathrm{n,elec}}$	3.92	3.92	μV
$v_{\mathrm{n,in}}$	2.56	46.70	μV

Finally, the amplifier should operate from a low supply-voltage to be compatible with the digital signal processing and a single-cell battery.

Table 3.2 summarizes the amplifier specifications.

3.3.2 Existing Neural Amplifiers

Most neural spike amplifiers implemented to date use a capacitively-coupled amplifier to remove the dc offset voltage and implement large impedances for low-frequency signals in a small area. Furthermore, the passband voltage gain of this amplifier is well-defined by

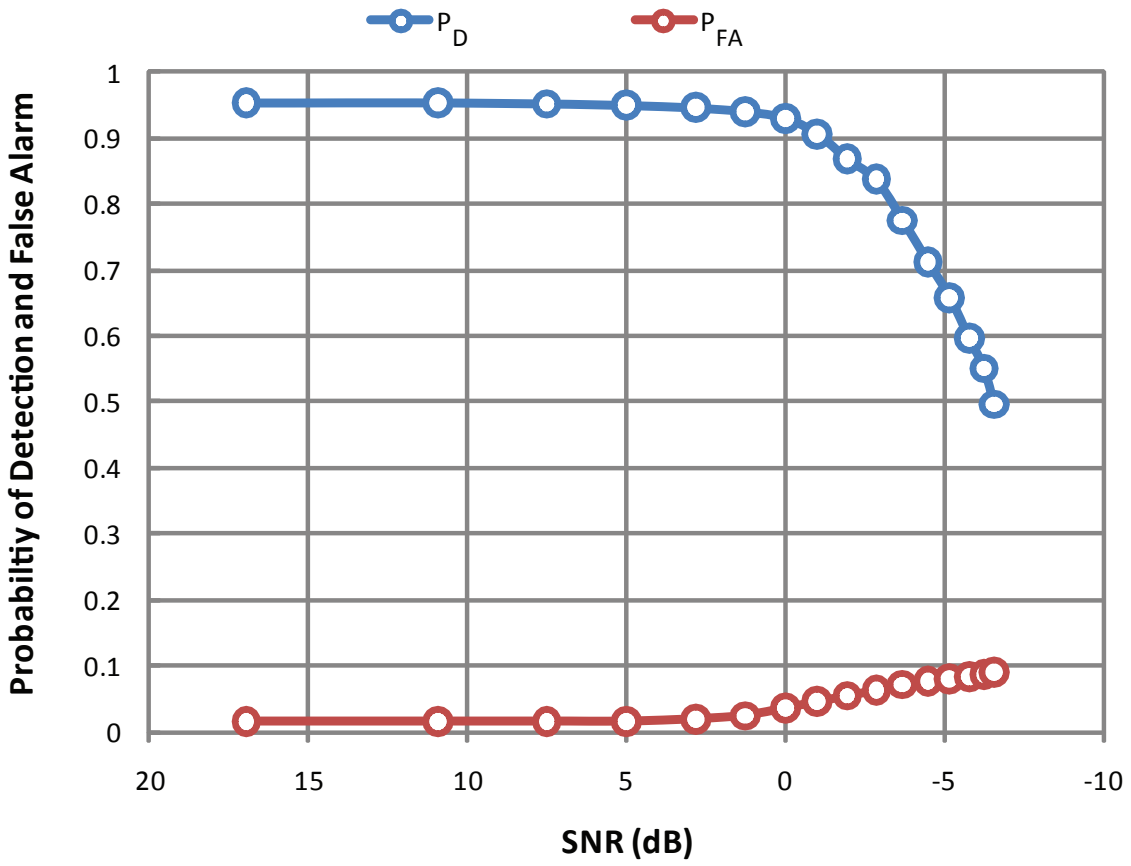

Figure 3.5: Required SNR of different spike detection algorithms. For a NEO detector, 5-dB is the minimum SNR for good detection and false-alarm probabilities. Based on data from [53].

Table 3.2: Spike Amplifier Specifications

Input Referred Offset Voltage	<1 mV	Input Referred Noise Voltage	<2.5 μV
Filter Corner (High-Pass)	100 Hz	Filter Corner (Low-Pass)	10 kHz
Voltage Gain (Differential)	80 dB	Voltage Gain (Common-Mode)	−52 dB
Output Swing	1 V_{p-p}	Output Load	2 pF

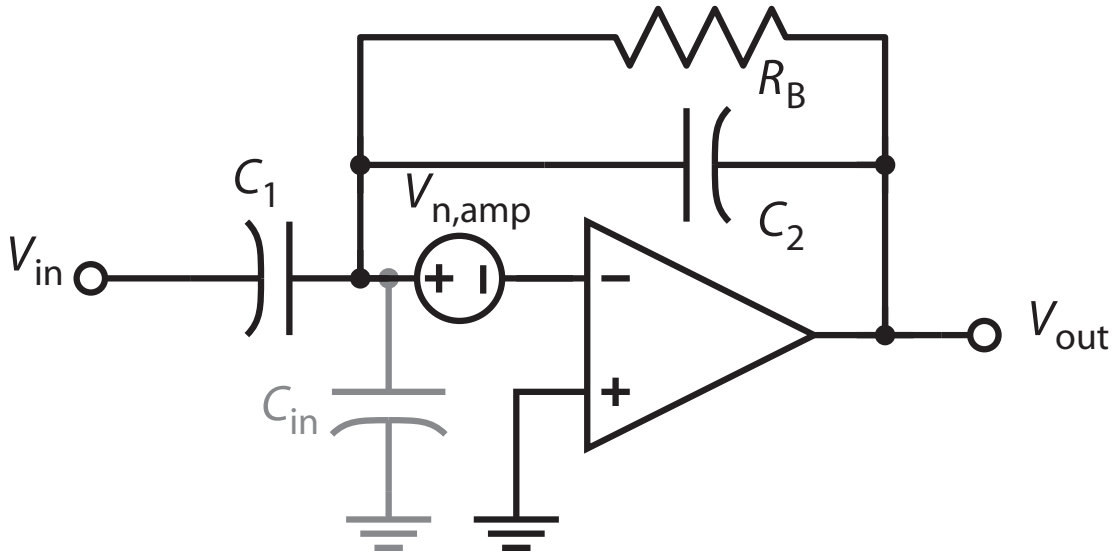

Figure 3.6: Capacitively Coupled Amplifier. $v_{n,amp}$ is the noise of the amplifier.

the ratio of the two capacitors C_1 and C_2. The high-pass corner frequency is $1/(R \cdot C_1)$. When the resistance is implemented by a pseudo-resistor [54], a giga-ohm resistance can be implemented in a small area. However it is non-linear, so cannot be used for the signal path if distortion is a concern. The low-pass pole is constructed by the finite bandwidth of the amplifier. A capacitor-based feedback network also achieves high input impedance with physically small components – using resistors for a similar input impedance would introduce a large parasitic capacitance which would increase power consumption and area. As the electrode capacitance is typically greater than 200 pF, which is much larger than C_1 used in typical integrated circuit implementations, the frequency response from a neuron to the output of the amplifier is determined solely by the circuit elements shown in Figure 3.6. Examples of this amplifier architecture can be found in [41, 49, 55–59]. Alternate approaches such as chopper stabilization [60] and dc stabilization [61] are also used but the most power-efficient designs to date use ac-coupled amplifiers. AC coupled amplifiers also completely isolate the dc offset from the amplifier (though transient disturbances still mandate the use of a mechanism to reset the low frequency (i.e. slow settling) bias network). As the dc offset is difficult to predict, ac coupling is preferred for this work. Furthermore, dc isolation between the biological tissue and electronics is required to avoid chemical interaction, even under fault conditions.

3.3.3 Amplifier Design

3.3.3.1 Noise Contributions

The input-referred noise contributions for the amplifier configuration shown in Figure 3.6 are summarized by Eqs. 3.2 and 3.3. As the high-pass corner f_l is pushed lower, the noise contributed by the feedback resistor R_B (used for dc biasing) contributes less noise. However the corner cannot be made arbitrarily low as the achievable resistance has an upper limit, and unwanted signals such as ac mains and LFP will also be amplified if the corner is made too low. The value of R_B is fixed after the capacitor size C_1 and f_l are chosen, and R_B does not appear in Eq. 3.3 explicitly. As f_l is essentially determined by application requirements, the key variable for optimization is C_1 in the familiar kT/C form.

$$\overline{v_{\mathrm{in,amp}}^2} = \left[\frac{1 + s \cdot R_\mathrm{B} \cdot (C_1 + C_2 + C_\mathrm{in})}{1 + s \cdot R_\mathrm{B} \cdot C_1}\right]^2 \cdot v_{\mathrm{n,amp}}^2 \tag{3.2}$$

$$\overline{v_{\mathrm{in,res}}} = \sqrt{\frac{2}{\pi}\frac{1}{A_\mathrm{V}} f_\mathrm{l}} \cdot \sqrt{\frac{kT}{C_1}} \cdot \frac{1}{f} \cdot \Delta f \tag{3.3}$$

As the gain is given by Eq. 3.4, and the electrode noise is injected at the input of the circuit, the electrode noise has the same frequency response as the amplifier itself. The resistance of the electrode therefore sets the minimum achievable noise (along with the background noise of thousands of neurons near the electrode). As frequency dependence of electrodes is often negligible for neural spike amplifiers, modeling the electrode as a single capacitor is appropriate, and there is no additional frequency dependence (Eq 3.5).

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{s \cdot R_\mathrm{B} \cdot C_1}{1 + s \cdot R_\mathrm{B} \cdot C_2} \tag{3.4}$$

$$\overline{v_{\mathrm{elec}}^2} = 4 \cdot k \cdot T \cdot R_\mathrm{E} \cdot |H(s)|^2 \Delta f \tag{3.5}$$

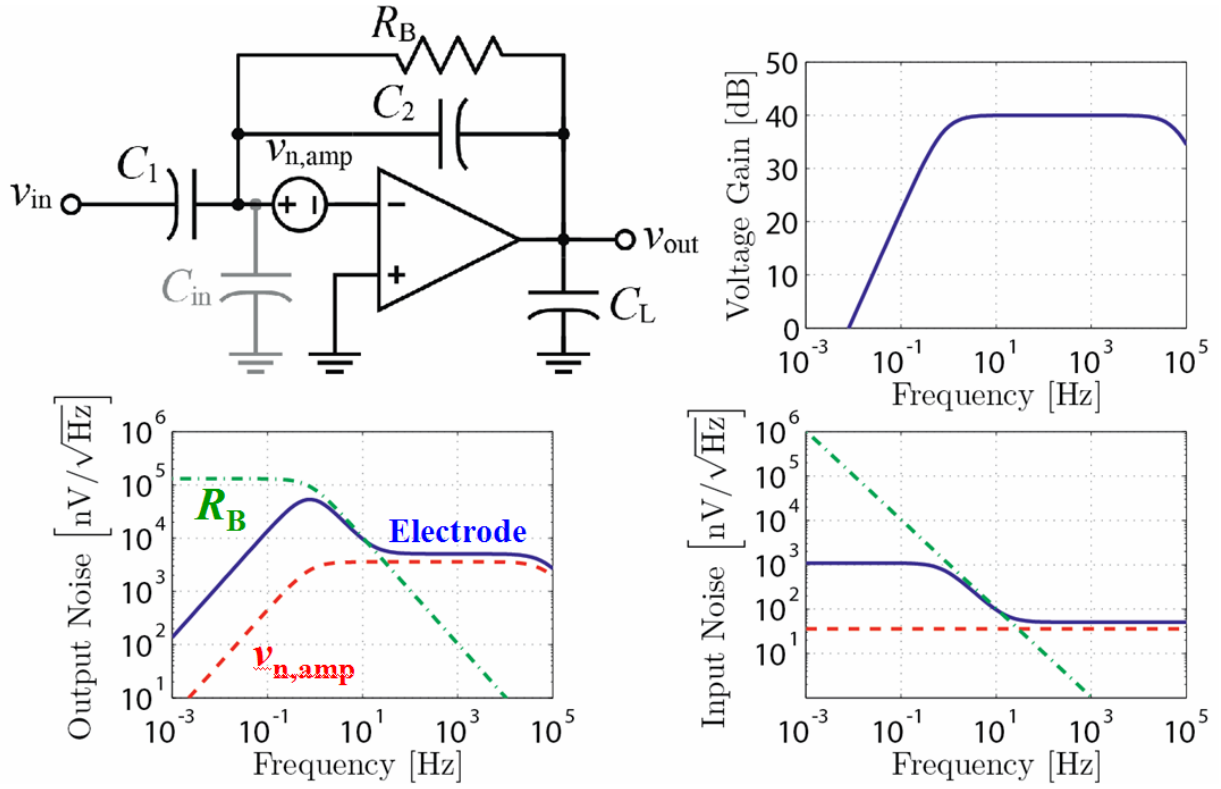

Figure 3.7: Noise contributions for an ac-coupled neural spike amplifier.

The noise equations given are for a single-ended circuit, with two times the noise power present in differential implementations. Fig. 3.7 shows that the electrode noise is the dominant noise source. As previously mentioned, the noise generated by the electrode is outside the control of the circuit designer. In Fig. 3.7, a more sophisticated electrode model has been used which causes the noise peaking around 1 Hz. The next largest noise source is the amplifier contribution, and will be the focus of subsequent optimization.

3.3.4 Transistor-Level Design

We now consider the design of the operational amplifier. The lowest power design to-date uses a folded-cascode operational transconductance amplifier (OTA) [59], and is shown in Figure 3.8. Approximate dc bias voltages are also shown, assuming that the input and output common-mode voltages are equal (as is the case when a resistor from output to input is used to set the common mode input level).

Because of the tall stack of transistors required to implement an op-amp, operation at low-voltage is difficult. We will look at an open-loop amplifier with a minimum number of stacked transistors to compare the power efficiency.

The total noise power, when input referred, can be given approximately by $v^2_{\mathrm{n,M1}} \cdot (1 + g_{\mathrm{m5}}/g_{\mathrm{m1}})$. Using Equations A.11 and A.12, the total noise can be plotted as a function of $V_{\mathrm{ds,sat}}$ and hence supply voltage (Figure 3.9). For the minimum VDD calculation, it is assumed that 0.5 V is sufficient for the other transistors, and $V_{\mathrm{ds,sat}}$ is allocated to M_5.

Now we see that the source of this degraded Op Amp performance at low voltage is due to limited headroom which leads to poor r_{ds}. Although we have not shown it here, stability concerns also limit the $V_{\mathrm{ds,sat}}$ that can be applied. Referring to Figure 3.9, we see when $V_{\mathrm{ds,sat}}$ is limited to 200 mV, the noise is limited to 1.4× the noise power of the input device. With a higher supply voltage, it is possible to reduce the noise contribution of the non-input devices. As shown in Figure 3.8, a degeneration resistor R_1 is used to reduce the transconductance of the current source M_5. Requiring a high supply voltage is in opposition to low power operation, so a different approach is needed.

To obviate this low supply-voltage issue, the Op Amp can be replaced with cascaded low gain stages, as shown in Figure 3.10. The analysis in Section 3.3.3 is still valid, with C_2 set to zero. The total noise and power in this design is dominated by the first stage even for moderate (greater than 5 V/V) values of voltage gain. However, in an Op Amp circuit, negative feedback is used to stabilize the voltage gain.

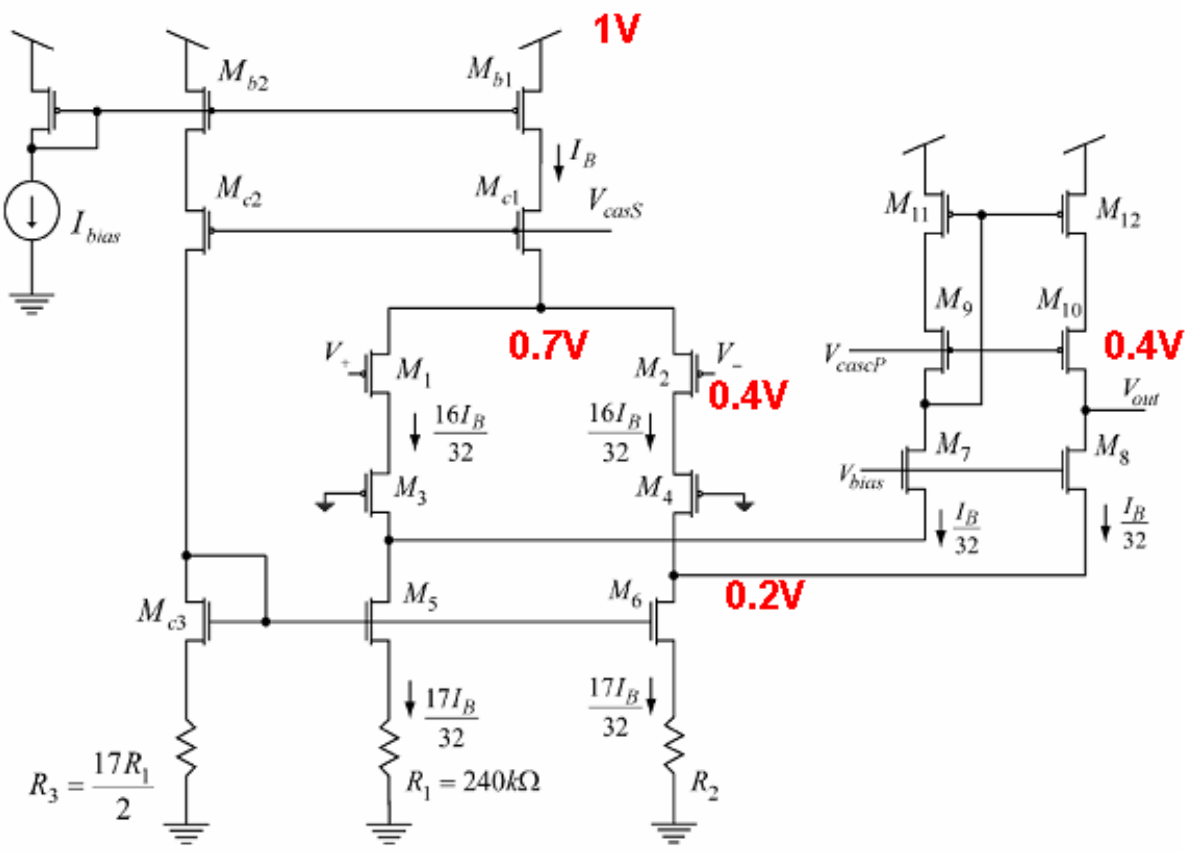

Figure 3.8: Amplifier used by [59], showing approximate bias voltages *if* operating at a 1 V supply.

To maintain a stable gain with our cascaded structure, calibration circuitry will be used. Because we aim to use a deep-submicron CMOS process and only need to track the relatively slow changes in dc gain, digital calibration will consume low area and power that what would otherwise be required to achieve similar noise performance with a single-stage design. To summarize, we will implement a cascaded amplifier structure capable of low-voltage operation, which takes advantage of the power and area-efficient digital processing available in deep-submicron digital CMOS.

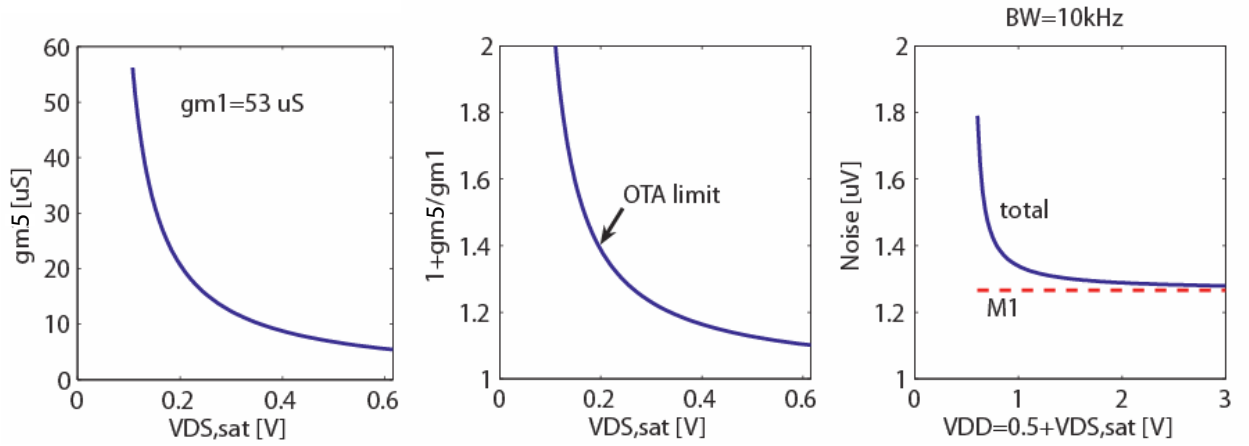

Figure 3.9: Relative noise from M_5 as a function of $V_{\mathrm{ds,sat}}$ and VDD.

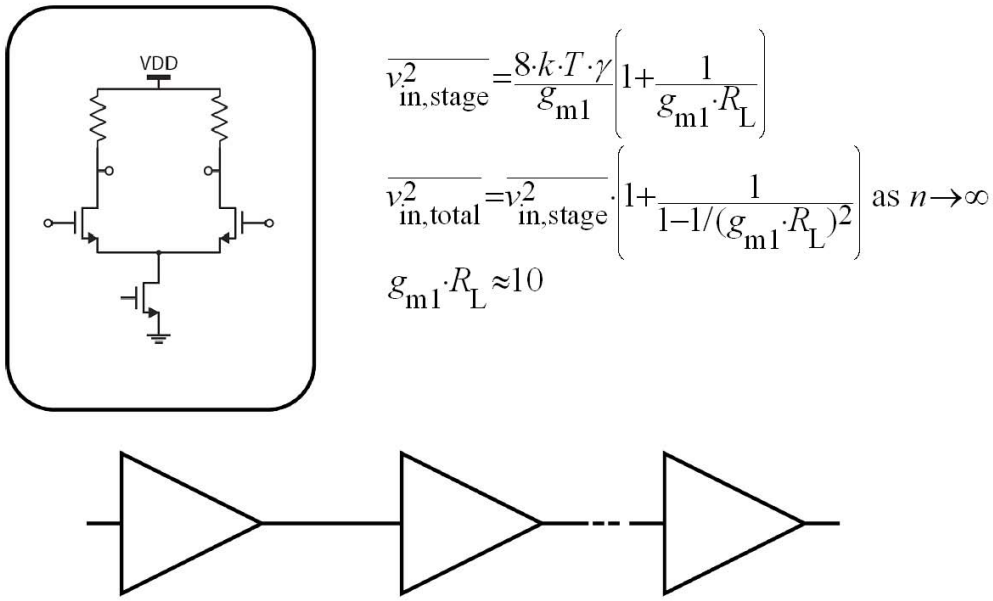

Figure 3.10: Cascaded Amplifiers to replace using Op Amps, when using a low supply voltage.

3.3.5 Optimization

For a given application, we now wish to determine the power and area of the preamplifier circuit. Capacitor C_1 and amplifier current I_D are the key specifications of the design and we will use these to determine the required amplifier power and area. First we will show the optimum bias current, and corresponding noise, for the amplifier with a given capacitor size. A key technique for low power designs is to bias the input devices in sub-threshold to maximize the $g_\mathrm{m}/I_\mathrm{D}$ ratio. As we also wish to minimize input capacitance, it is expected that biasing near the edge of sub-threshold, i.e. $IC \approx 0.1$, should give the lowest power. With this operating point in mind, the minimum supply current and corresponding noise can be derived.

3.3.5.1 Optimal Bias Current

For a fixed inversion coefficient (IC) or current density, we wish to sweep the bias current to minimize the total input referred noise $v^2_\mathrm{in,amp}$. The noise of the amplifier alone, with its input devices biased in sub-threshold, is given by Eq. 3.6. Since we have set a fixed IC, increasing the amplifier current to reduce noise also increases the input capacitance of the amplifier. The proportionality constant α relates supply current to input capacitance in the form $C_\mathrm{in} = \alpha I_\mathrm{D}$. Bias conditions and technology node affect the scaling between C_in and I_D, and it is desirable to have a small input capacitance for a given bias current, which implies α should be minimized.

$$v^2_\mathrm{n,amp} = 4 \cdot k \cdot T \cdot \gamma \cdot K_\mathrm{AMP} \cdot \frac{V_\mathrm{T}}{\kappa} \cdot \frac{1}{I_\mathrm{D}} \cdot \Delta f \tag{3.6}$$

where K_AMP is a proportionality constant relating the noise of the amplifier to the noise of a single transistor. K_AMP is essentially a function of the amplifier biasing and architecture used.

With the input capacitance C_in expressed as $\alpha \cdot I_\mathrm{D}$, the input referred noise is:

$$v_{\mathrm{n,in}}^2 = \left(\frac{C_1 + C_2 + \alpha \cdot I_{\mathrm{D}}}{C_1}\right)^2 \cdot v_{\mathrm{n,amp}}^2 \tag{3.7}$$

In order to minimize the $v_{\mathrm{n,in}}$ noise, C_1 must be made as large as possible. However this consumes a large amount of chip area, so we must trade-off capacitor area versus noise. Suppose we choose a value for C_1 based on an area requirement, and then determine the minimum achievable noise. The noise is large whenever I_{D} is small (Eq. 3.6) or large (Eq. 3.7), which implies that there is an optimum value somewhere in between.

Substituting Eq. 3.6 into Eq. 3.7 and solving for the minimum value of $v_{\mathrm{n,in}}$ w.r.t I_{D} yields the result that the minimum noise is obtained when I_{D} is equal to $(C_1 + C_2)/\alpha$. The optimum input capacitance is $C_1 + C_2$, independent of α and bias conditions. This is a key observation [1]: *the noise is minimized when the input capacitance C_{in} of the amplifier is equal to the sum of the source capacitors C_1 and C_2.* This observation allows us to quickly choose the input capacitance C_{in} from C_1 and vice versa. The input-referred noise-voltage is also two times the amplifier noise-voltage at the minimum. We have assumed the electrode impedance is much smaller than that of C_1; if it is not then its effect should be included.

The required current can be obtained by first solving Eq. 3.8 for W, given that IC was chosen for high $g_{\mathrm{m}}/I_{\mathrm{D}}$ efficiency and transistor length L was chosen to meet dc gain requirements. The current is finally given by Eq. 3.9

$$\begin{aligned} W &= \frac{I_{\mathrm{D}} \cdot \kappa \cdot L}{2 \cdot \mu \cdot C_{\mathrm{ox}} \cdot V_{\mathrm{T}}^2} \\ &= \frac{C_1 + C_2}{C_{\mathrm{ox}}} \cdot \frac{1}{1 - \kappa + IC} \cdot \frac{1}{L} \end{aligned} \tag{3.8}$$

$$\begin{aligned} I_{\mathrm{D}} &= I_{\mathrm{S}} \cdot IC \\ &= \frac{2}{\kappa} \cdot \mu \cdot C_{\mathrm{ox}} \cdot V_{\mathrm{T}}^2 \cdot \frac{W}{L} \cdot IC \end{aligned} \tag{3.9}$$

[1] It is also not surprising when considering the maximum power transfer theorem.

Using the EKV equations in Appendix A, we can derive an expression for α valid in the weak inversion region:

$$\alpha = \frac{\kappa \cdot L^2}{2 \cdot \mu \cdot V_{\mathrm{T}}^2} \cdot \left(1 + \frac{1-\kappa}{IC}\right) \tag{3.10}$$

Repeating the previous derivation using this definition of α yields the minimum noise in Eq. 3.11:

$$\begin{aligned} v_{\mathrm{in,min}}^2 &= 16 \cdot \frac{1 + C_2/C_1}{C_1} \cdot \frac{V_{\mathrm{T}}}{\kappa} \cdot k \cdot T \cdot \gamma \cdot K_{\mathrm{AMP}} \cdot \alpha \cdot \Delta f \\ &= 8 \cdot \frac{1 + C_2/C_1}{C_1} \cdot \frac{\gamma \cdot K_{\mathrm{AMP}} \cdot q \cdot L^2}{\mu} \cdot \left(1 + \frac{1-\kappa}{IC}\right) \cdot \Delta f \end{aligned} \tag{3.11}$$

Several key points are worth noting regarding Eq. 3.11. The ratio C_2/C_1 is $1/A_{\mathrm{V}}$ where A_{V} is the amplifier's passband gain. The gain should then be maximized to reduce the input noise. Minimizing L also minimizes the noise, but L may be constrained by other considerations such output resistance r_{ds}. NMOS devices should be used to maximize mobility μ, as long as flicker noise is not a concern. Biasing should maximize IC also, up to the edge of weak inversion. Finally, after the other parameters have been set, we can select C_1 to meet the noise requirement (for a given process technology). While scaled technologies provide shorter channel lengths, the increased gate leakage current processes after the 0.18 μm node introduces other issues which must be solved before using these technologies for the input device. Using the thick-oxide I/O transistors (which are typically available in all processes) for the analog portion of the design and thin-oxide transistors for the digital design provides a good compromise.

Figure 3.11 shows the input referred noise for 3 different capacitor sizes. Three curves are shown, for different values of C_1. The larger C_1 is, the smaller the minimum noise that can be achieved, although it requires a larger current I_{D} to reach the minimum noise. For the neural spike amplifier described in this dissertation, an input noise of 2.5 μV has been

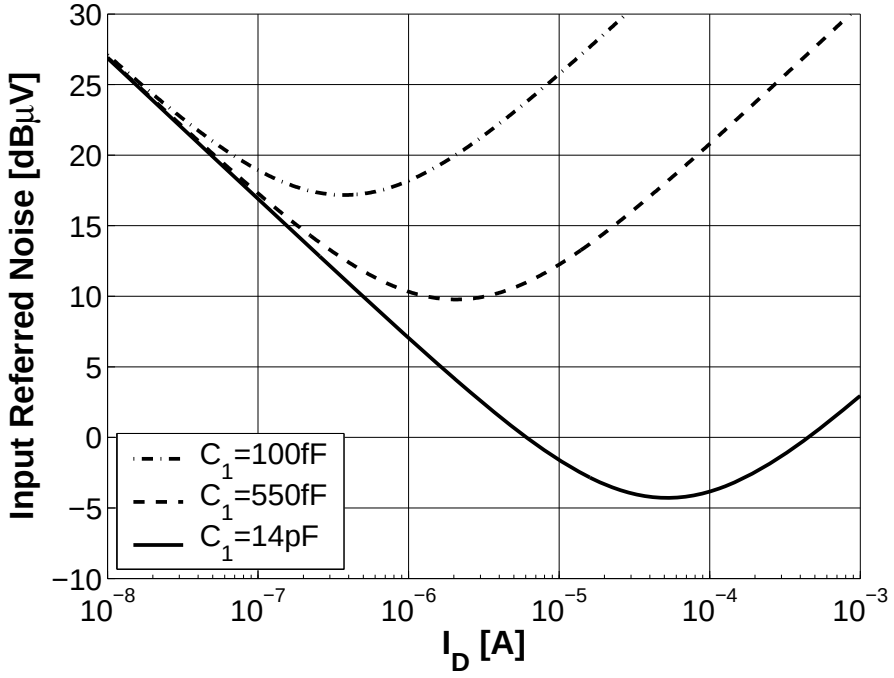

Figure 3.11: Plot of Input Referred Noise for different C_1 values. For a Neural Spike Amplifier, 14 pF is required. (With IC=0.1 and an NMOS input.)

selected. Since we know for a given capacitor size for C_1, although we do not yet know what size that is, the amplifier noise must be half of the input referred amount, or 1.25 μV. For a differential circuit (which requires $\sqrt{2}$ lower noise per device), and assuming a 50/50 split of thermal and flicker noise (again $\sqrt{2}$ for uncorrelated noise sources), the thermal noise of M_1 ($v_{\mathrm{n,th-M1}}$) must be less than 0.625 μV. The drain current can then be calculated from $I_{\mathrm{D}} = 4kT\gamma U_{\mathrm{T}}/\kappa/v^2_{\mathrm{n,th-M1}}$, which is approximately 6.6 μA. This current requires a device W/L of 100μm/0.5μm, which presents around 75 fF of gate capacitance. However, in the next section, we will see that flicker noise imposes a larger device area, and thus sets the size of C_1.

An alternate approach to optimize the amplifier would be to maximize the SNR. When the input capacitance of the amplifier is comparable to the coupling capacitor C_1, a reduction in gain is expected. If the gain is reduced substantially, then the noise contributions of subsequent stages have more effect on the SNR measured at the output. An amplifier with 60 dB dc gain and 10 MHz unity-gain frequency (UGF) configured for 40 dB passband gain (i.e. C_1/C_2=100) is analyzed with the aid of Fig. 3.12.

The size of the amplifier is scaled as we move along the x-axis from small input capacitance, which has corresponding low power and high noise, to large input capacitance which in turn has low noise and high power. The gain is measured at 10 kHz which is the upper passband frequency. Even when the amplifier has small input capacitance, a gain error of ~1 dB results from finite dc gain. As the input capacitance increases, the feedback factor of the circuit is reduced, and the gain also is reduced, with a sharp decline as C_{in} approaches C_1. The noise behavior has been previously described, and reaches a minimum around $C_{\mathrm{in}} = C_1$. The SNR is also measured, and peaks when the input referred noise is minimized. Provided that the amplifier gain is high enough to suppress the noise of subsequent stages, the reduced gain does not affect the optimization, and the tradeoff between area through C_{in} and power still applies.

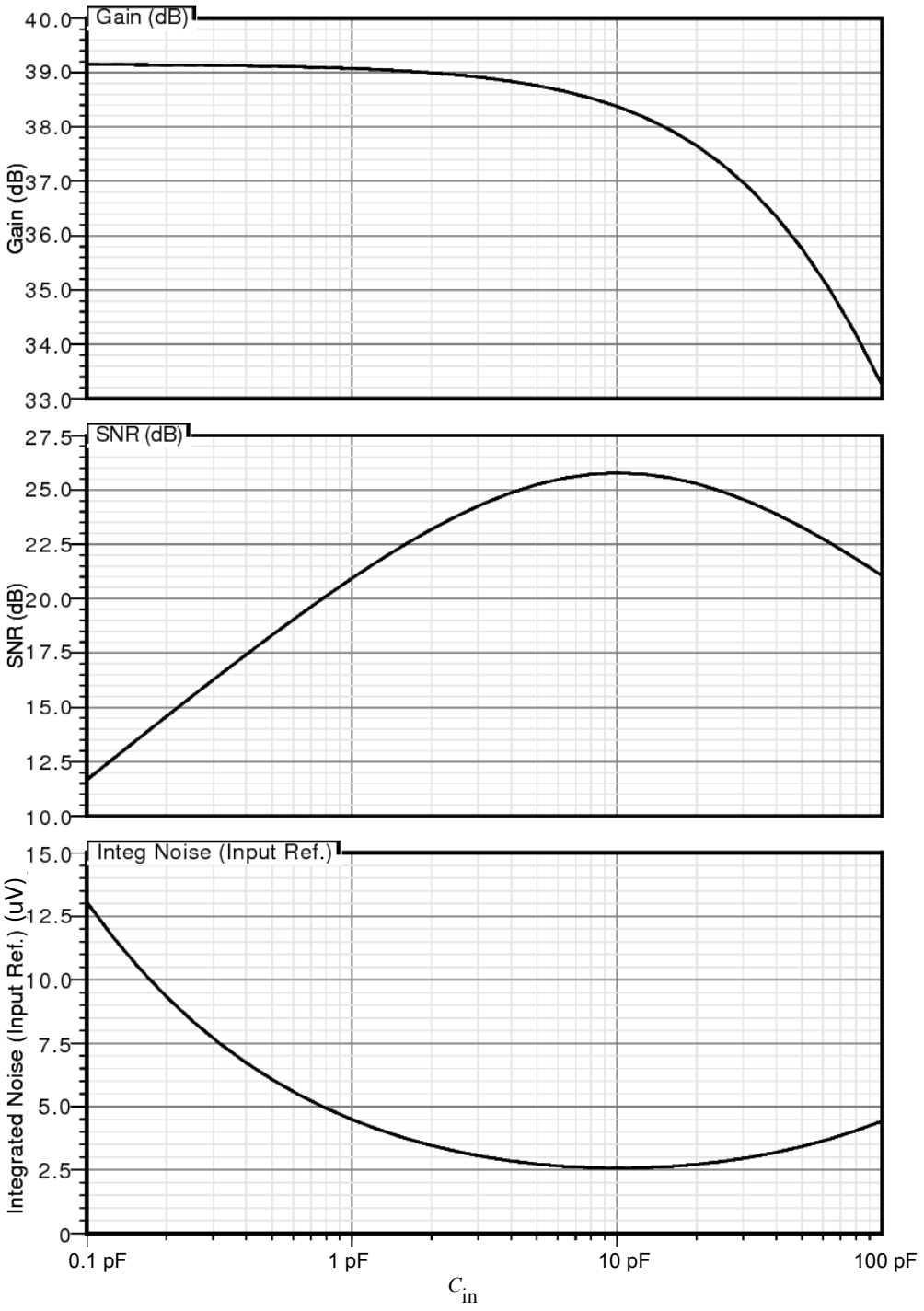

Figure 3.12: Amplifier gain, noise, and SNR as function of input capacitance C_{in}.

3.3.6 Accounting for Flicker Noise

So far, meeting the noise specification has focused on thermal noise. For low-frequency designs, flicker noise is also significant. Given the optimized design based on thermal noise constraints, we calculate the additional noise from flicker noise. If gate-referred noise power spectral density for a MOSFET in saturation is modeled by

$$S_{\text{flicker}}(f) = \frac{K_{\text{F0}}}{W \cdot L} \cdot \frac{1}{f}\cdot, \tag{3.12}$$

we see that gate area $W \cdot L$ must be chosen to reduce the noise, and a process-dependent constant K_{F0} determines the absolute value.

The input-referred noise over a band from f_1 to f_2 is given by

$$\begin{aligned} v^2_{\text{n,in}-\text{flicker}} &= \left(\frac{C_1 + C_{\text{in}}}{C_1}\right)^2 \cdot \frac{K_{\text{F0}}}{W \cdot L} \cdot \ln(f_2/f_1) & (3.13)\\ &\approx \left(\frac{C_1 + C_{\text{ox}} \cdot (1-\kappa) \cdot W \cdot L}{C_1}\right)^2 \cdot \frac{K_{\text{F0}}}{W \cdot L} \cdot \ln(f_2/f_1) & (3.14) \end{aligned}$$

The value of W that minimizes the input noise is

$$W = \frac{C_1}{C_{\text{ox}} \cdot (1-\kappa) \cdot L} \tag{3.15}$$

This leads to a C_{in} that in turn sets the value for C_1. For the TSMC 65-nm process used for the amplifier in Chapter 5, flicker noise is high, and requires a gate area of 1874 μm^2 for 0.625 μV of flicker noise over a band from 100 Hz to 10 kHz. The minimum capacitance C_1 for this amplifier is 1.4 pF. Using a large C_1 will result in lower input referred noise at the expense of more area. With an input device with W/L of 3750μm/0.5μm, and assuming the input devices are 2/3 of the total area, we can calculate the space available for C_1 assuming we can place capacitors over active devices. In this case, we can make C_1 as large as 3.1 pF before it is larger than the amplifier itself (assuming 1fF/μm^2 for the capacitors). Since C_1 is

greater than the minimum required, we could reduce the device and capacitor sizes, or allow thermal noise to increase in exchange for reduced supply current. However, this still serves as an initial estimate, and leaves some margin for modeling inaccuracy and other effects.

Finally, we discuss the choice of channel length, since flicker noise only specifies a certain gate *area*. We can implement this area as a wide device with a short channel, or increase both width and length (compared to the thermal-noise-limited design). With a wide device (i.e. increased W only to increase area), IC is reduced, and for a large device (i.e. increased both W and L), IC is constant. In both cases the input capacitance C_{in} is larger than what we needed for thermal noise.

Rewriting the equation for input capacitance in weak inversion (A.8) as

$$C_{\text{g,weak}} = (IC + 1 - \kappa) \cdot C_{\text{ox}} \cdot W \cdot L \tag{3.16}$$

$$= \frac{\kappa \cdot I_{\text{D}}}{2 \cdot \mu \cdot U_{\text{T}}^2} \cdot L^2 + (1 - \kappa) \cdot C_{\text{ox}} \cdot W \cdot L \tag{3.17}$$

shows that increasing L affects both terms. For this reason, it is preferred to increase W. Either option is feasible though, as increasing W will also increase the source/drain area of the MOSFET which was not included in this analysis, and the second term dominates in 3.17.

3.3.7 Adjustable Biasing

The design given produces an amplifier that meets the requirements for a minimum amplitude signal. However, in a system with a large number of electrodes, many channels will have higher amplitudes and will exceed the SNR requirement. Reducing the current from the maximum value so that the no more than the required SNR of 5 dB is maintained will maximize the efficiency of the overall system while keeping the same detection accuracy.

The current can be reduced in two ways. One is to reduce the amplifier bias current. In this regime, which we call adjustable bias, the input capacitance remains roughly constant

(mostly due to gate-to-bulk capacitance) and amplifier noise $v_{\mathrm{n,amp}}$ increases with decreasing current. The second method is using several amplifiers connected in parallel to adjust the current (and noise). In both cases, the coupling capacitance C_1 is constant. In this second method, which we call parallel amplifiers, we are switching in additional amplifiers when needed, and the input capacitance increase as the current increases. The effect of increased current is stronger than the increased amplifier input capacitance, and the noise decreases. In comparison to the adjustable bias method, only the current changes, rather than device operating points such as IC. The results are shown in Fig. 3.13.

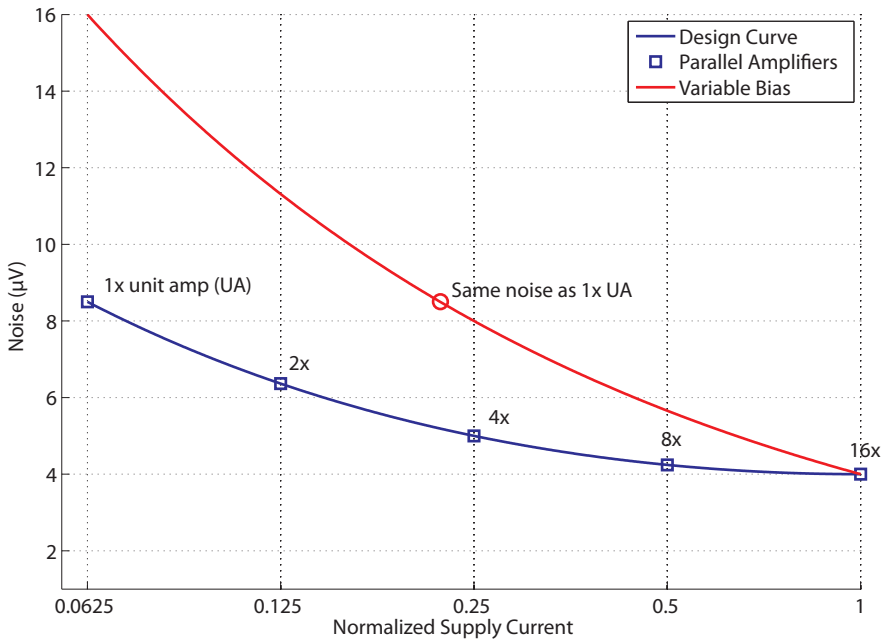

Figure 3.13: Adjustable bias with parallel amplifiers or variable bias current. Parallel amplifiers give a better noise-power tradeoff.

We will estimate the capacitance and current for an amplifier with a 10-kHz noise bandwidth and 2.5 μV noise. Beginning with a normalized current of 1, we set the bias to current to meet the noise requirement. This occurs with $\approx$1.4 pF for C_1 and I_{D} of 6.6 μA. As the current is reduced for variable biasing, the noise power is inversely proportional to the current,

and increases rapidly. In contrast for parallel amplifiers, we can use 16 unit amplifiers (UA) in parallel for the lowest noise. Decreasing the number of amplifiers in parallel decreases the current and input capacitance while increasing the noise. The plot shows the noise for a configuration with 1, 2, 4, 8 and 16 amplifiers in parallel. For all these configurations, the parallel amplifiers yield lower noise for the same supply current than the variable bias case. Using parallel amplifiers would allow an aggressive reduction in current. The drawback is the added wiring complexity and parasitics to connect these amplifiers, when needed. A combination of both techniques could also be used to achieve a wide range of operating points.

3.3.8 Summary of the Design Methodology

The design methodology can be summarized as follows:

1. Determine the required input-referred noise for the amplifier from system requirements.
2. Initially, for a minimum capacitor size, assume that the amplifier noise must be half the input-referred noise.
3. Assume flicker noise and thermal noise contributed equally, and determine the device current to meet thermal noise requirements. This also determines the input capacitance required for a thermal-noise limited design.
4. Calculate the gate area to meet the flicker noise requirement, and compare this to the area required for the thermal noise case.
5. Capacitor C_1 is determined by the larger of C_{in} based on thermal and flicker noise constraints.
6. If the fabrication process allows capacitors over active devices, compare the size of C_1 to the input devices, and increase C_1 until it covers the amplifier.

7. After the initial design is complete, refine the design with simulation.

3.4 Analog-to-Digital Converters

Low-power ADCs are a critical part of many applications, and as such, several examples exist in the literature that are suitable for use in a biosignal-data-acquisition systems.

Based on recent published work (such as [62–64]), ADCs with a figure-of-merit (FoM) on the order of 100 fJ/conv-step are readily achievable in the 8 to 12 bit range. Fig. 3.14 shows a survey of ADCs against different FoMs. As technology and circuit architectures improve, the FoM also improves. In the future, FoMs ten times lower may be commonplace [64]. Power increases linearly with speed, and exponentially with resolution (and often faster than 2^B). Area is primarily determined by the resolution.

If we assume FoM is determined by reviewing comparable ADCs with different resolutions and speeds, we can estimate the power of the ADC that meets our requirements. Our estimated power P_{ADC} is given by

$$P_{\mathrm{ADC}} = FoM \cdot F_{\mathrm{s}} \cdot 2^B \tag{3.18}$$

where B is the number of bits, and F_{s} is the sampling rate.

A low FoM indicates that little energy is expended for each conversion of a sample. Lines of constant FoM are also shown on Fig. 3.14. Several ADCs achieve a performance close to 100 fJ per conversion-step, with two close to 10 fJ/conv-step. We use 100 fJ/conv-step as our benchmark for estimating ADC power. ADC area can also be estimated from the survey (not shown). The most power efficient ADCs use the successive approximation register (SAR-ADC) architecture. Their high power-efficiency, moderate speed, and medium resolution match the needs of biosignal acquisition.

For an 8-bit, 24 kSa/s, 100-fJ/conv-step ADC, the power consumption is expected to be 614 nW; we will use this as our benchmark ADC power. Silicon area ranges from 0.021 to

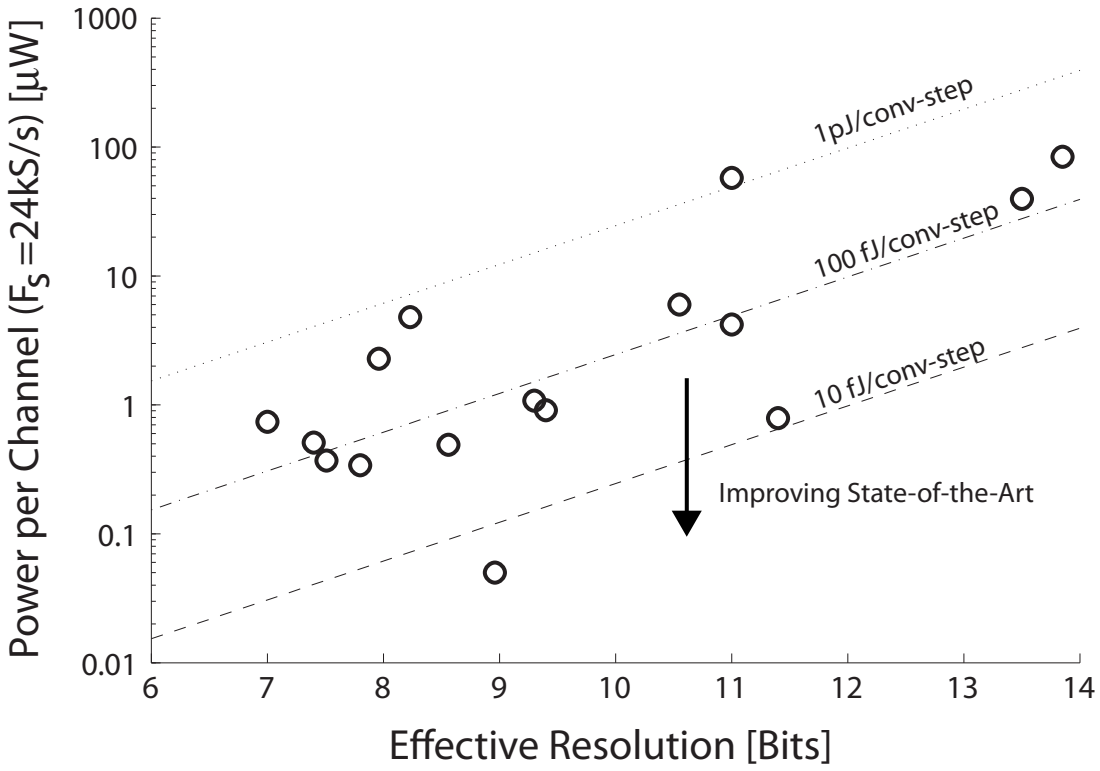

Figure 3.14: Plot comparing of recent low-power ADCs.

0.24 mm^2; for the estimates below we use 0.03 mm^2 (A_0) and 8 bits (B_0) as our baseline ADC area.

The area (A_{ADC}) of the ADC is modeled by Equation 3.19.

$$A_{\mathrm{ADC}} = A_0 \times 2^{2\cdot(B-B_0)} = 0.03 \times 2^{2\cdot(B-8)} \tag{3.19}$$

As we alluded to earlier, it may be beneficial to only use the ADC when a spike is detected. The system could be configured to wait for a spike to be detected, enable the ADC and digitize the spike, and enter a standby mode after the spike data is captured. That is, the ADC is operated with a duty cycle of less than 100%. With this mode of operation, the ADC digitizes the input with an average rate of r_{D} samples per second (which must be less than F_{s}). The exact role of r_{D} will be explained later in Section 3.5.4. A buffer is

also included to drive the ADC and store samples in an analog memory (See Section 3.5.3.3). The power of this buffer is given by P_{buf}. It is assumed that the standby power of the ADC is 10% of the ADC when sampling at full speed, and that the buffer does not dissipate power in standby mode. The power averaged over operational and standby modes is given by Eq. 3.20,

$$\begin{aligned} P_{\text{ADC,eff}} &= (0.9 \cdot FoM \cdot 2^B + P_{\text{buf}}/F_{\text{s}}) \cdot r_{\text{D}} \\ &\quad +0.1 \cdot FoM \cdot 2^B \cdot F_{\text{s}} \end{aligned} \tag{3.20}$$

where F_{s} is the sampling rate (24 kSa/s) and r_{D} is the detection rate in samples/s, defined later in Eq. 3.22.

We also assume that interleaving the ADC with up to 64 channels has little impact on the overall performance.

In this section we have described a mathematical model for the power and area of the ADC as a function of resolution B and sampling rate F_{s}. We will use this model to help determine the system power.

3.5 Spike Detection

One question that arises when optimizing spike-detection hardware is whether spike detection should be performed in the analog or digital domain.

Many existing systems perform spike detection in the analog domain (e.g., [35, 43, 49, 65–67]), while others choose to perform spike detection in the digital domain (e.g., [68–70]). The assumption is that analog spike detection is more power-efficient since the ADC would only need to run when spikes are detected, whereas in digital detection the ADC must be running constantly, since detection occurs only after sampling (Fig. 3.15). However, performing computations in the digital domain has the advantage that digital-design techniques, such as voltage scaling and interleaving, can be employed that are not trivial in the analog domain.

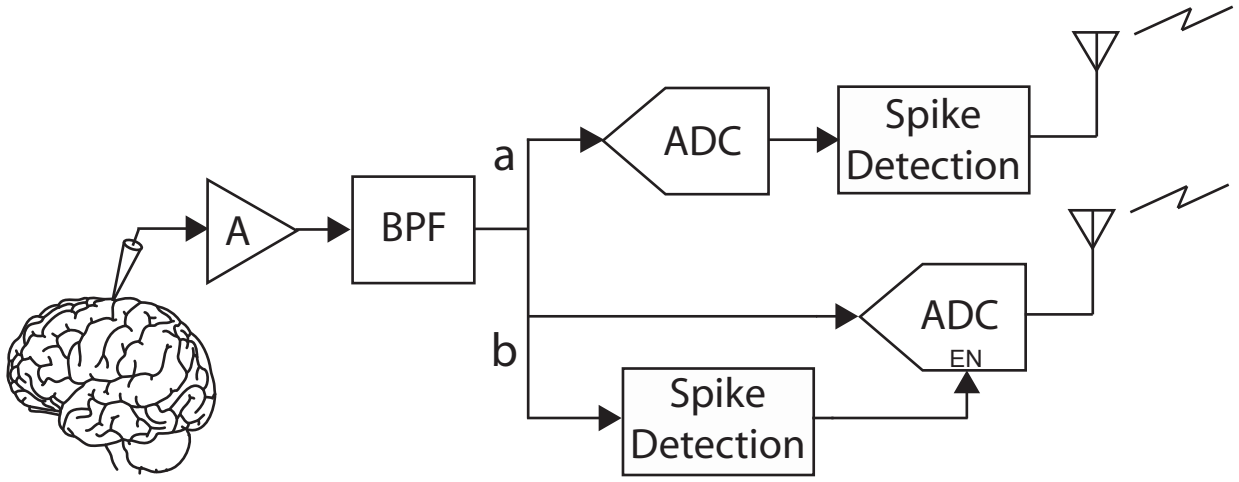

Figure 3.15: Block diagram for (a) digital spike detection and (b) analog spike detection.

In this section, we attempt to determine whether analog or digital spike detection is more efficient, with respect to both power and area. Circuits are simulated in 65-nm bulk CMOS, with thick oxide 250-nm transistors used for some parts of the analog designs.

3.5.1 Spike-Detection Algorithms

Because many algorithms for spike detection exist, we chose to analyze a couple of algorithms with different computational complexities. Implementing the algorithms in the analog or digital domain will result in different power and area for the complete system. The preferred algorithm will have lower overall power and/or area.

The two chosen algorithms are absolute-value thresholding and nonlinear energy operator. In absolute-value thresholding, a threshold is applied to the absolute value of the waveform $x(n)$ [71]. In the nonlinear energy operator (NEO) method [71–74], a threshold is applied to the NEO ψ:

$$\psi[x(n)] = x^2(n) - x(n+1) \cdot x(n-1). \tag{3.21}$$

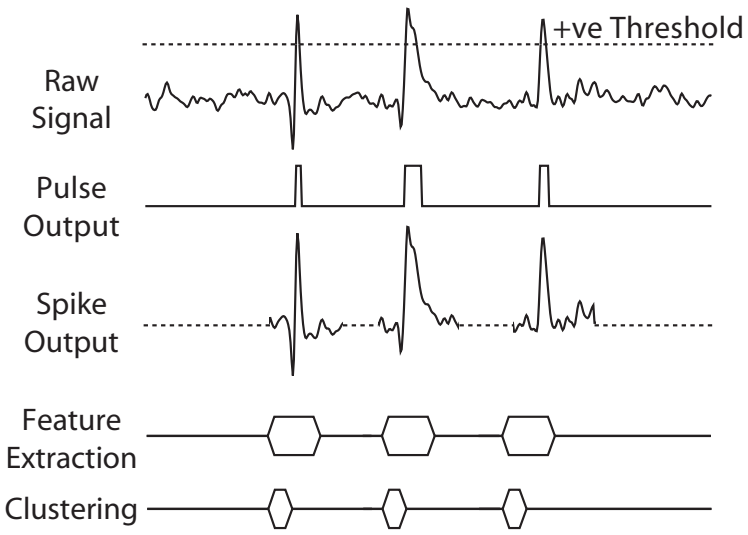

Figure 3.16: Outputs of spike detector for each mode of operation. In `Pulse Output` mode, a pulse is outputted each time the waveform crosses the threshold. In `Spike Output` mode, spike samples are outputted for optional feature extraction and clustering.

3.5.2 Modes of Operation

We also chose to analyze each algorithm for two different modes of operation (Fig. 3.16). In `Pulse Output` mode, the spike detector only outputs a pulse when the signal ($|x|$ or ψ) crosses the threshold. Most published analog spike detectors operate in this mode. It is the simplest mode, since it requires neither memory nor an ADC (in the case of analog spike detection). However, in applications that require single-unit activity, spike sorting must be performed following spike detection. If only spike times are outputted, then the spike shapes are lost, making subsequent spike sorting impossible. Therefore, the second mode of operation (`Spike Output`) that we analyzed is transmitting 1-ms-worth of waveform samples before the threshold crossing (the "spike preamble") and 2-ms-worth of waveform samples after the threshold crossing. Although 3 ms is much longer than a typical spike, this provides a sufficient number of samples for subsequent alignment.

3.5.3 Analog Spike Detection

Before we consider different implementations of the analog spike detectors, we will briefly consider the signal levels. With this information, we are able to specify tolerable degradation due to the analog circuit (such as noise and offsets). As shown in Fig. 3.15, the detector circuitry is placed after amplification. Typical extracellular spike amplitudes are in the 50 to 500 μV range, with worst-case noise of 10 μV. With a preamp voltage gain of 100 V/V, the noise at the detector and ADC input is 1 mV. The analog detector must keep its own electronic noise and offset voltages below this level. While this is achievable with modest power, we compare the analog and digital implementations to determine which is optimal. We will only describe the dominant power and area contributors; clock power, for instance, was found to be negligible at the operation frequency of neural spike recording. The supply voltage assumed is fixed at 1 V for the analog portions. Area calculations are based on total active MOSFET area ($W \times L$) and capacitors with a specific capacitance of 1 fF/μm^2.

3.5.3.1 Absolute-Value Threshold Detector

Absolute-value thresholding can be performed with a clocked comparator and a switched-capacitor (SC) difference circuit. The primary error is the comparator offset voltage. Smaller offsets can be achieved by increasing the device area at the cost of power dissipation. The other significant error source is charge injection from the switches.

The comparator shown in Fig. 3.17 draws approximately 0.176 μA from a 1-V supply. The gate area of the input devices is 10 μm^2, yielding a random offset of 10.8 mV (in 65-nm CMOS). Assuming an ADC full-scale of 500 mV, this offset would result in a 10% error in the desired threshold. However this could be overcome by circuit techniques such as auto-zeroing, or by increasing the device area further. The area is dominated by the sampling capacitor C_1 (500 fF), which leads to an area estimate of 500 μm^2.

A reference circuit is also required. The supply current of approximately 0.2 μA would

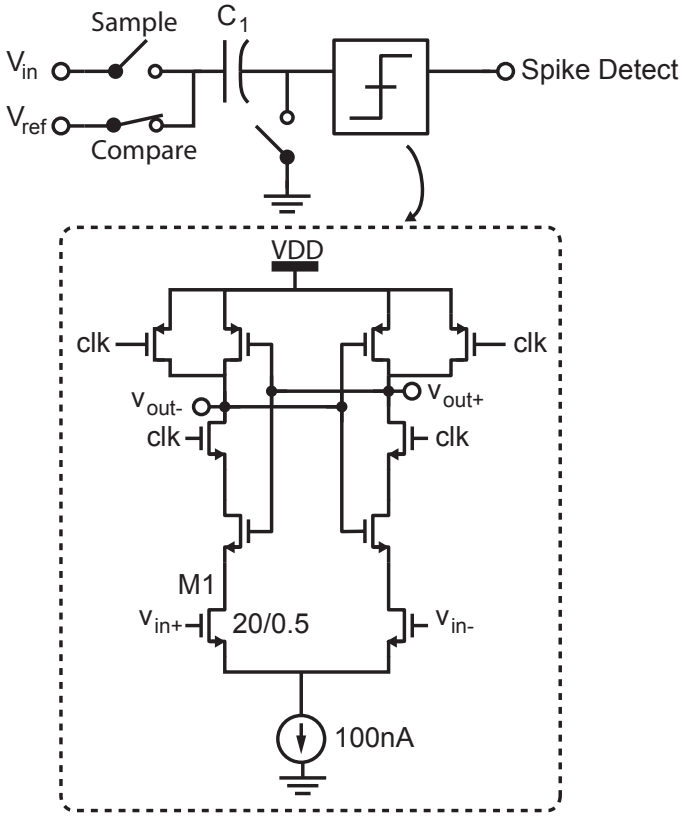

Figure 3.17: Schematic for a low-power dynamic comparator.

drive the input capacitance of the comparator in its comparison phase. For absolute-value thresholding, two comparators are used to detect positive and negative crossings. The total power of an analog spike detector is $2\times0.18 + 0.2 = 0.54$ μW.

3.5.3.2 Analog Nonlinear Energy Operator (NEO) Detector

Because analog continuous-time differentiation is prone to being noisy, its time constant is sensitive to process variation, and for a more direct comparison to the digital implementation, we implement the discrete-time version of the NEO as shown in Eq. 3.21.

In order to implement Eq. 3.21 we require an analog multiplier and analog memory. The algorithm can be implemented with the circuit shown in Fig. 3.18. After the first half of

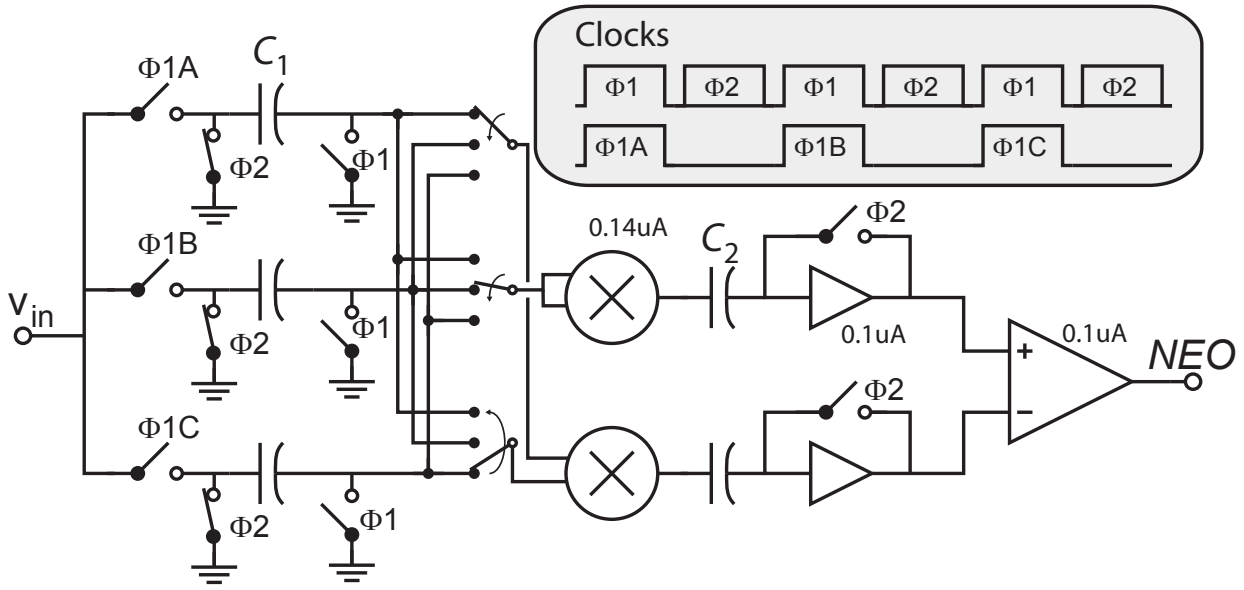

Figure 3.18: Implementation of the Non-Linear Energy Operator algorithm in the discrete–time analog domain.

the clock period ($\Phi 1$), a new sample is buffered in the C_1 array. During $\Phi 2$, the products $x^2(n)$ and $x(n-1) \cdot x(n+1)$ are computed from the corresponding capacitor voltages. A commutator after the capacitor array routes the correct sample to the multipliers, and the routing is updated each clock cycle. The products are stored on C_2. Auto-zeroing of the multipliers is achieved during $\Phi 2$, with C_2 implementing Output Offset Storage (OOS). This allows smaller devices to be used in the multipliers to save area.

Because the linear input range of a typical Gilbert multiplier is on the order of ± 50 mV, only limited gain can be applied to the signal. Hence thermal noise of the multiplier may cause excessive degradation of the SNR. Simulations show that 140-nA tail current for the multiplier is sufficient for 150-μV noise (1.5 μV at the preamplifier input).

The power of the combined circuitry (2 multipliers and 3 amplifiers, plus comparator) is $2 \times 0.14 + 3 \times 0.1 + 0.38 = 0.96$ μA. The analog NEO circuit requires a total of 5 pF of capacitance, so the total area is approximately 5000 μm^2.

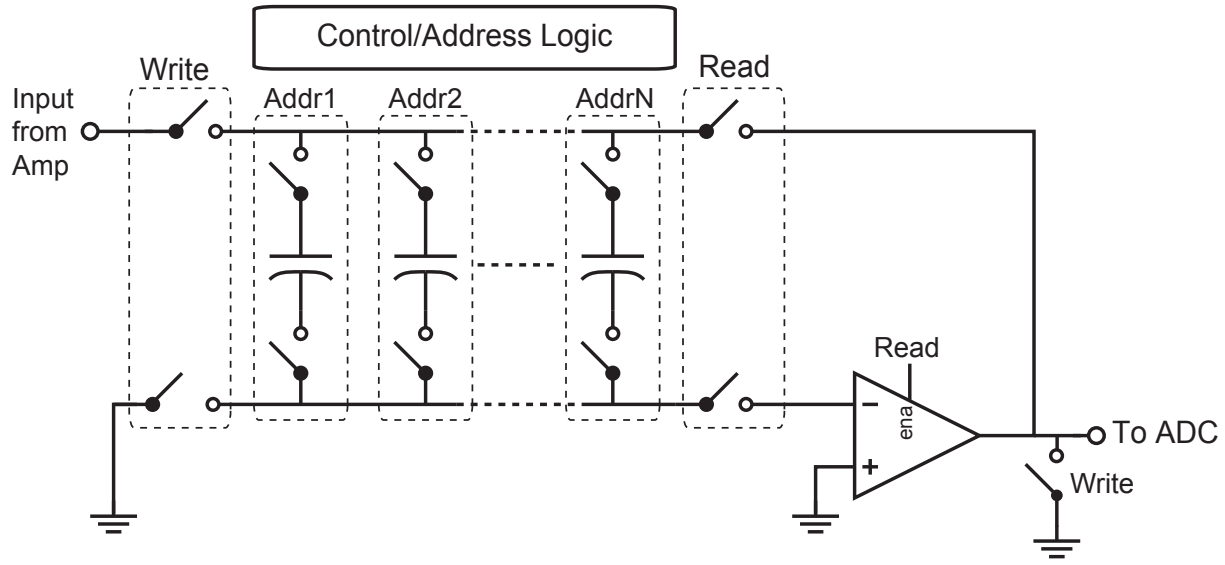

Figure 3.19: Implementation of analog memory for storing the signal before a spike has been detected.

3.5.3.3 Analog Memory

In some applications, it is advantageous to retain the samples before the spike detect event. This is easily done in a digital implementation, but it somewhat difficult to do in an analog implementation. One solution, by Anelli [75], is shown in Fig. 3.19. In other applications that do not require the spike preamble, e.g. `Ppulse Output` mode, we can ignore this power and area.

Setting the storage capacitance (C_1) as 100 fF meets kT/C (thermal) noise requirements. The total area is computed as 24 samples $\times$ 100 fF $\times$ 1 fF/μm^2 $\times$ 2 (for a differential implementation) yielding a total area of around 4800 μm^2.

After a spike is detected, the memory must be read by the ADC. A buffer, in the form of a flip-around track-and-hold, provides good linearity. To estimate the power of this amplifier, we assume a two-stage OTA. Behavioral simulations show that $g_{\mathrm{m1}} = 0.564$ μS is required

to meet the settling time requirement. The total opamp current, assuming g_{m2}=3g_{m1} is then $8 \times (g_{m1}+g_{m2}) \times V_T/\kappa = 180$ nA. Our estimate for the total amplifier is 1.5× the area of the compensation capacitors (2 pF), for a total area of 3000 μm^2.

3.5.4 Effect of SNR and Firing Rate on Analog Detection Power

Since SNR and firing rates can vary significantly across neural recordings, it is important to check the validity of this analysis for a wide range of SNRs and firing rates. For example, if the rate of detection increases as SNR decreases due to false alarms, then the power of analog spike detection would also increase due to more frequent use of the ADC. If so, there could be a range of SNRs and firing rates for which digital detection is more efficient. Therefore, we estimated the variation in power consumption for the analog detection hardware with SNR and firing rate.

We generated data with the neural signal simulator used in [45] for SNRs ranging from about 15 dB to -10 dB. We then performed spike detection using both algorithms, using the automatic threshold calculation techniques described in [45], and calculated the probability of detection (P_D) and the probability of false alarm (P_{FA}) at each SNR. These rates were then used to calculate the detection rate r_D in samples per second for each SNR using Eq. 3.22:

$$r_D = \max\{r_N \cdot l \cdot P_D + (F_S - r_N \cdot l) \cdot P_{FA}, F_s\}, \tag{3.22}$$

where r_N is the firing rate of the neurons (which can be the sum of firing rates of multiple neurons) in spikes per second and l is the length of a spike in samples per second. r_D can be thought of as the number of samples that the ADC must convert/quantize per second. Note that the maximum value that r_D can take is F_s. This equation was used in Eq. 3.20 to calculate the power of the ADC.

Figure 3.20 shows the variation in power of the analog implementation of NEO with

`Spike Output` mode. The detection rate, and therefore the power, remain constant across SNRs until around -5 dB, when the number of detections (power) begins to decrease. This is due to the adaptive nature of the threshold, which is based on a multiple of the mean of the NEO, and which, therefore, increases as the noise increases. Figure 3.20 also shows that the power increases linearly with the firing rate, due to the linear increase in detection rate with firing rate, with a maximum variation in power of about 600 nW. (Similar results, not shown, were obtained for the absolute-value method.) To simplify the analysis, we use an operating point of 1.3-dB SNR and 100-Hz firing rate when presenting the results in Section 3.5.6.

3.5.5 Digital Spike Detection

In order to obtain power and area estimates for the digital implementations of the spike-detection algorithms, both the absolute-value threshold and the NEO detection methods (`Spike Output` and `Pulse Output` modes as explained earlier) were implemented in the Matlab/Simulink-based design environment. Each of the above algorithm was implemented with 2, 4, 8, 16 and 32 channel data-stream interleaving to determine a power-area efficient implementation. The RTL was auto-generated from the Simulink model using the Synplify DSP blockset. Power and area estimates were then obtained from the synthesis reports for these designs when synthesized with DC compiler from Synposys. Simulated neural data was input to RTL simulations to obtain switching activity estimates for the design. These estimates were then annotated into the synthesis flow to obtain power estimates for the digital spike-detection module.

Based on technology evaluation results for our design in 65-nm bulk CMOS process, we chose to operate the circuits at a reduced supply voltage of 0.4 V. Since standard-cell libraries are characterized for the nominal supply voltage (1 V), we specified a higher clock frequency for synthesis in order to account for the increase in delay due to supply voltage scaling. We

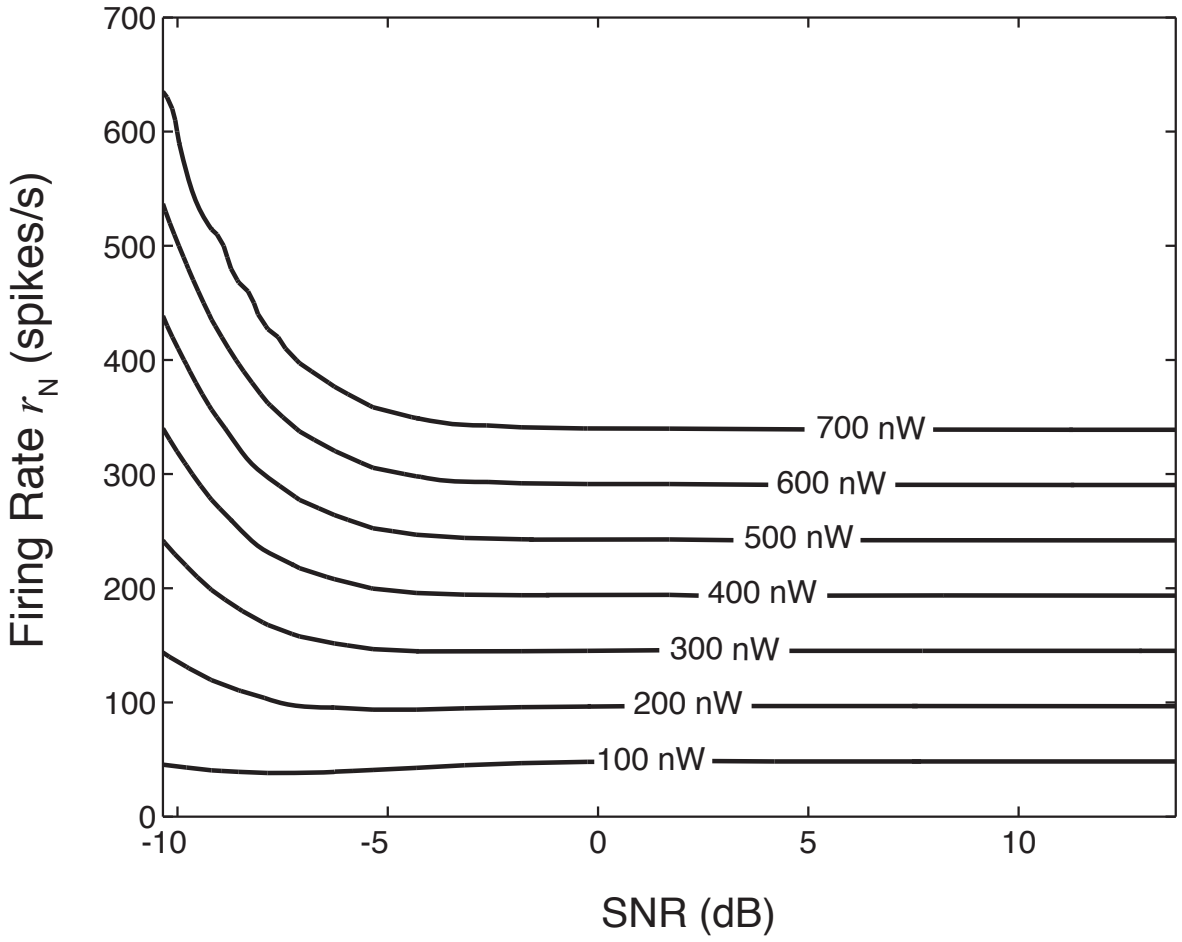

Figure 3.20: Variation in power of analog NEO spike detection, `Spike Output` mode, due to changes in SNR and neuronal firing rates. (Note: Firing rate can represent the sum of firing rates from multiple neurons.) The variation was calculated by subtracting the minimum power from each value. The power remains constant across SNRs until about -5 dB, at which point it begins to decrease, and the power increases steadily with firing rates, until saturation, when the ADC is operating at its maximum 24 kSa/s.

also evaluated the reduction in leakage power due to supply voltage scaling for basic gates. The switching power and leakage power numbers obtained from synthesis were thus scaled down to their corresponding values at 0.4 V to make comparisons for power consumption.

Figures 3.21 and 3.22 show the area and power per channel versus the number of channels interleaved. Interleaving usually increases the power due to increased switching activity of logic and a similar number of registers switching at a faster rate. However, if the supply voltage is scaled, savings in the leakage power of logic and the increase in switching power are comparable. Thus, the total power consumed per channel versus degree of interleaving has a global minimum. From the above results we find 8-channel interleaving to be a power-area efficient implementation for the detection algorithms considered. Also we observe that area and power for `Spike Output` mode are significantly higher than those for `Pulse Output` mode. This is due to the additional logic and memory required to provide the detected spikes (with preamble) as the output. It should be noted that we use a register-based memory for our design to guarantee functional operation at 0.4 V. However, use of a custom low-voltage memory would reduce the power difference between `Pulse Output` and `Spike Output` mode implementations. We find that variation in SNR and firing rate does not cause significant variations in the power consumed by the DSP. This result is expected, due to two major reasons: a) SNR and firing rate do not affect the ADC power for digital detection b) SNR and firing rate only affect the switching power of a portion of the DSP, which does not cause a significant change in the total power of the DSP scaled to 0.4 V. Hence, we expect the results of the above analysis to be valid for a wide range SNR and firing rates.

3.5.6 Results

Figure 3.23 shows the power per channel and area per channel for each algorithm and output mode. The first row of plots corresponds to power per channel, and the second row of plots correspond to area per channel. Solid red (blue) lines correspond to the total power/area per channel of analog (digital) detection, including the power and area of the ADC. The solid red (blue) line can be decomposed into the power/area of detection alone, indicated by the dashed red (blue) line, and the ADC power when operating at the maximum rate (F_{s}), indicated by the dashed black line.

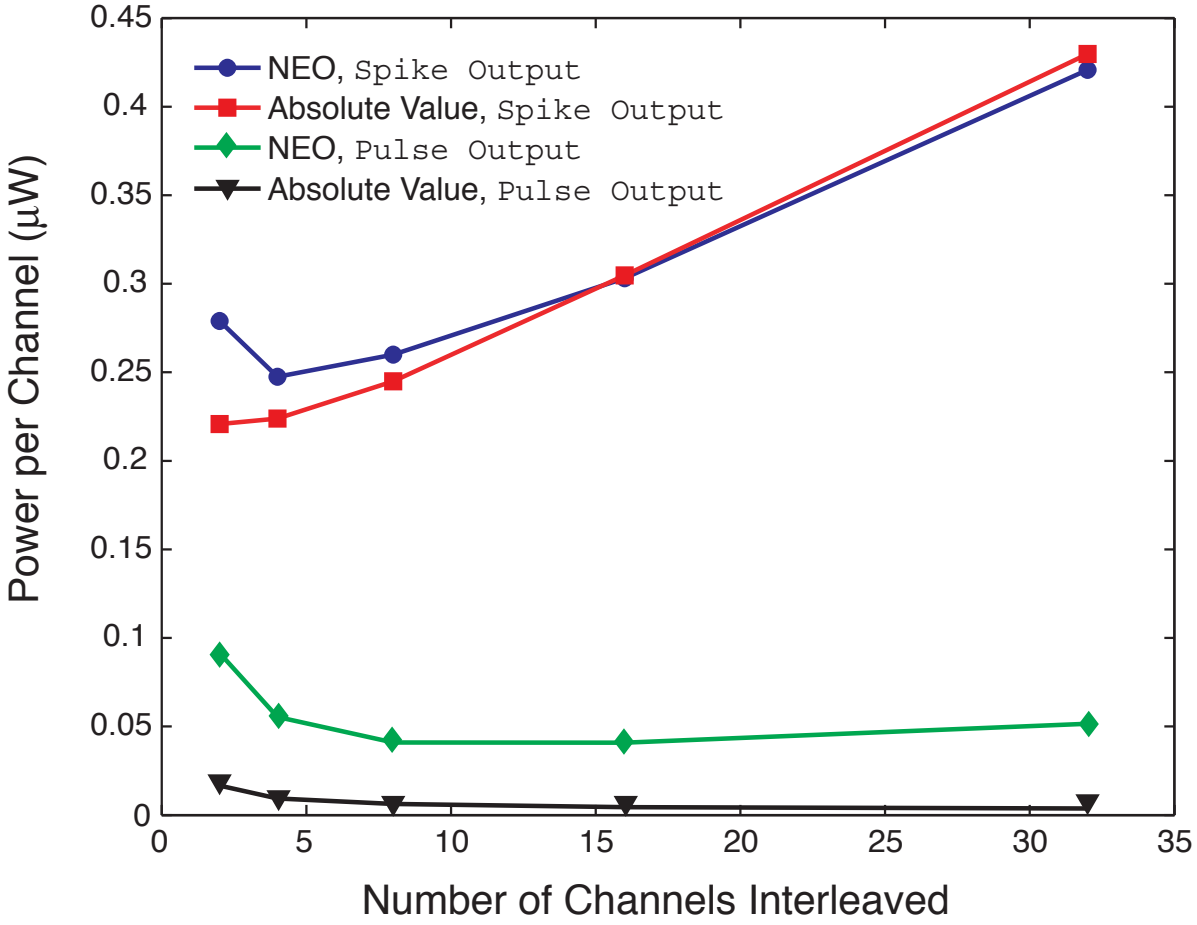

Figure 3.21: Power estimates obtained from Synopsys for NEO, `Spike Output` mode. The total power (P_{total}) is divided into switching power ($P_{\text{switching}}$) and leakage power (P_{leakage}).

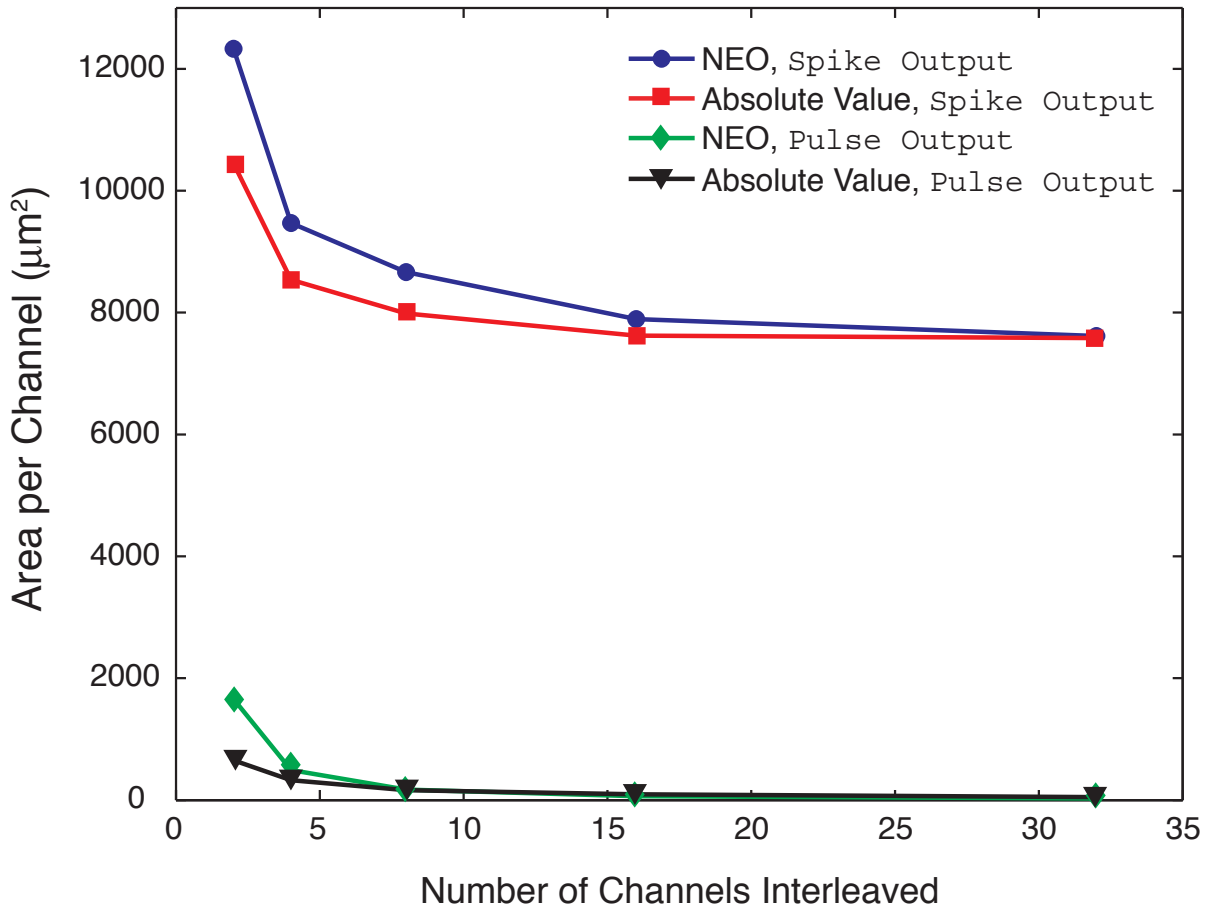

Figure 3.22: Area estimates for NEO, `Spike Output` mode, obtained from Synopsys as a function of the number of channels interleaved.

The power for analog detection in the pulse mode is constant across bit resolution since there is no need for the ADC. In the case of the `Spike Output` mode, the power for analog detection does increase with bit resolution, since the ADC power increases. However, the increase in analog detection power due to the ADC is less than that for digital detection. This is because the ADC is only active for a limited time in case of analog detection. As for the area tradeoff, there is a crossover between analog and digital implementations in the case of `Pulse Output` mode. At higher bit resolution, the area for analog implementation is less since the area of the ADC dominates. However for the `Spike Output` mode, the area of digital detection is less than that of the analog detection.

From these plots we can conclude that digital spike detection is better for resolutions of up to 8 or 9 bits, depending on the algorithm used. Until this point, the ADC power is small. Hence the saving in ADC power achieved by analog detection does not outweigh the lower power cost of the DSP implementation. However, since the power and area of the ADC scale exponentially with bit resolution, the ADC starts to dominate at higher resolution. The analog detection, therefore, is more power-efficient in this domain.

3.5.7 Summary of Spike Detection

We have compared the power consumption and the area of analog and digital spike detection. We demonstrated that power is not a strong function of SNR or firing rates; thus, the results shown for the operating point 1.3 dB, 100 Hz are valid across a wide range of SNRs and firing rates. We also showed that the tradeoff between digital and analog detection is a strong function of the bit resolution. For lower resolutions, digital implementations are more efficient, whereas for higher resolutions, analog implementations are more efficient. Therefore, the choice of whether to implement hardware spike detection in the analog or digital domain is dependent on the desired resolution.

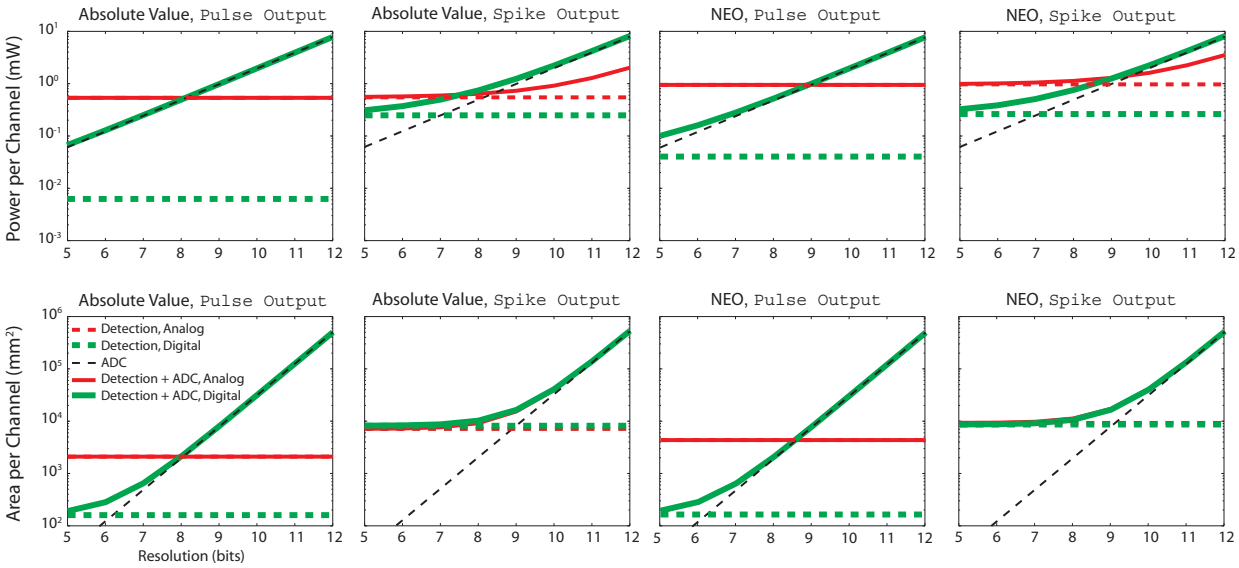

Figure 3.23: Power per channel and area per channel for each algorithm and output mode. The first row of plots corresponds to power per channel, and the second row of plots corresponds to area per channel. Solid red (blue) lines correspond to the total power/area per channel of analog (digital) detection, including the power/area of the ADC. The solid red (blue) line can be decomposed into the power/area of detection alone, indicated by the dashed red (blue) line, and the ADC power/area when operating at the maximum rate (F_s), indicated by the dashed black line. For each algorithm and output mode, a crossover point exists between analog and digital, indicating that for lower resolutions digital spike detection is more power/area- efficient, while for higher resolutions analog spike detection is more power-efficient.

3.6 Digital Signal Processing

There are two aspects worthy of discussion regarding digital signal processing. First, how much signal processing should be performed before transmitting data off-chip? Second, how is the chosen system implemented efficiently?

To address the question of on-chip processing, several options are considered. Three chosen options for spike processing are (1) detection and alignment only, (2) feature extraction, and (3) clustering. Implementation of detection has been previously discussed, but we revisit it in the context of the whole system.

In conventional circuits, either the signal is passed directly to the transmitter, or only detected spikes are sent. Feature extraction and clustering can be performed off chip. Clearly, passing the raw data to the transmitter requires the least hardware, has the highest fidelity since the waveform is uncompressed, but also has the highest datarate. Detection and alignment requires additional hardware with the advantage that the datarate can be reduced. Feature extraction and clustering requires even more hardware to implement, and substantially reduce the datarate. Lossy compression such as adaptive differential pulse code modulation (ADPCM) can also reduce the data rate of the raw data down to 65%, while maintaining a correlation with the original signal of 99.9% [76, 77].

Because both feature extraction and clustering greatly reduce the amount of data required to represent a spike, and hence can reduce the power required for the transmitter since fewer bits are transmitted.

Table 3.3, which is based on data published by Karkare [78], shows the power consumption per channel for each of the different options, with their corresponding data rate of processed signals per channel. Immense reductions in data rate can be achieved, which in turn eases the load on the transmitter. The amount of data compression is dependent on the spike firing rate. Table 3.3 assumes a firing rate of 100 Hz, a 10-bit ADC, and 48 samples/spike [79].

Table 3.3: DSP Power for Different Levels of Processing

	μW/chan	kbps/chan
Raw Data	0	300
Spike Detect	1	48
Feature Extraction	4	10
Clustering	8	0.4

With these power estimates, we will be able to calculate the required data rate that the transmitter must support. With the datarate, we will also be able to estimate the power of the transmitter.

3.7 Wireless Transmitter

When designing a wireless telemetry link, one must consider the physical limitations of the transmitter, the channel (i.e., the medium between the transmitted and the receiver), and the receiver. A full wireless-link budget, which calculates the required transmitter power level, starts at the output of the transmitter and ends with demodulated data from the receiver. An expression for the required transmitted power level as a function of the critical physical limitations involved, is given by

$$P_{\mathrm{TX}} = \frac{(2 \cdot k \cdot T \cdot R_{\mathrm{S}}) \cdot NF \cdot SNR \cdot BW}{PL \cdot RFM \cdot G_{\mathrm{RX}} \cdot G_{\mathrm{TX}}} \quad . \tag{3.23}$$

$$FSPL = \left(\frac{4 \cdot \pi \cdot d}{\lambda}\right)^2 \tag{3.24}$$

For a 10 m link at 2.4 GHz operation, the path loss is 66dB. Assuming a receiver sensitivity of -70 dBm and antenna gain of 10 dB, the required output power from the headstage is -14 dBm.

The first group of terms represents the noise generated by a 50-ohm resistor (R_{S}) in a matched RF system, which takes into account the impedance at the antenna. The second term is the noise figure NF, which is the ratio of the noise at the output of the receiver to the noise contribution due to a 50-ohm resistor passed through the receiver. The third term is the SNR required for decoding the digital data with a bit-error rate of less than 10^{-6}. Although this error rate may seem high, conventional coding strategies can be used to reduce the error rate to a level required for a given application. The last term of the numerator is the bandwidth of the communication channel. The terms of the denominator involve critical components of the communication channel. The path loss PL, represents the reduction of transmitted power as a function of distance from the transmitter. Rayleigh-fading margin RFM takes into account the changes in received power due to the constructive or destructive overlap of signals arriving from multiple paths (i.e., multipath interference). The transmitter antenna gain G_{TX} takes into account the impact of the antenna design on its ability to efficiently transmit power to the channel. Similarly, the receiver antenna gain G_{RX} takes into account the impact of the design of the receiver antenna on its ability to receive power from the channel. This term will also include gain achieved through multiple-input-multiple-out (MIMO) strategies when used, although we expect that in this application there will be only a single input (i.e., SIMO) [80].

Ultimately, the multipath issue imposes a limit on the maximum data rate that can be achieved for a given communication channel. A transmitted signal may take multiple paths to the receiver. As a result, there is a spread in arrival times of a given transmitted signal at the receiver. The symbol length is the name given to the duration of time used to transit a unique representation of a bit pattern. The symbol time must be significantly greater than the spread in arrival times. Typically, a factor of 10 is considered to be acceptable. The delay spread (i.e. the rms value of arrival times at the receiver) of a typical room is approximately 20 ns [81, 82]. Given the 10× design rule-of-thumb, the symbol length must be at least 200 ns (i.e., 5×10^6 symbols/s). By encoding two bits into each symbol, the maximum data rate

is 10 Mbps at the cost of transmitter complexity.

Numerical values for each component of the link budget are given in Table 3.4 (which shows Eq. 3.23 in log form) and are either directly calculated or taken from literature. The result of all of this analysis, is that the minimum power delivered by the transmitter to the channel must be at least 12.6 μW (-19 dBm). Of course, this value is dependent on the selection of the modulation scheme, the design of the individual components, and the specific needs of the application. Implementing an efficient transmitter to deliver the required output power is still an active research topic.

A 2.4 GHz transmitter has been previously published by Zolfaghari for use in an IEEE 802.11b or Bluetooth radio with a power dissipation of 12 mW, and an output power of 0 dBm. The maximum datarate of this system is 11 Mbps, which is close to our desired maximum data rate of 10 Mbps. Since our required output power is much lower, the bias current of the transmitter can be substantially reduced. Compared to the original 0 dBm output power, an output power of -19 dBm would require a reduced output voltage swing that is $10^{19/20}$ times smaller. This raises the possibility that the power amplifier could be removed, and the antenna driven directly from the mixer. Without a heavy load on the mixer, the switches, which load the local oscillator (LO) buffers could also be reduced. The power of the transmitter is composed of 7.5 mW in the power amplifier, 3.75 mW in the LO buffers, and 0.75 mW in the upconversion mixer. We estimate that the power dissipation could be reduced by removing the power amplifier and decreasing the power of the LO buffers by 50%, to yield a transmitter power of 2.6 mW. Based on this estimation, we claim that 3 mW for the transmitter is feasible.

3.8 Conclusion

The optimization of several key blocks of a wireless telemetry system has been described. The optimization of amplifier noise and area, as a function of capacitor size and bias current, has

Table 3.4: Simple RF Link Budget

Noise Power	-174	dBm	50 Ω at 300 K per Hz
Receiver NF	8	dB	Conservative Estimate
Required SNR	10	dB	BPSK @ BER=10^{-6}
Bandwidth BW	70	dB	10 MHz
Path Loss PL	62	dB	$n = 2$, $d = 10$m, $f = 2.4$GHz
Rayleigh Fading RFM	20	dB	
RX Diversity	-15	dB	
Required P_{TX}	-19	dBm	

been described. State-of-the-art ADCs have been reviewed, and shown to have sufficiently low enough power dissipation to be compatible with a low-power biosignal telemetry system. Digital signal processing, at both the algorithm and circuit level have been discussed, along with strategies for minimizing power (i.e. selecting an NEO algorithm to maintain reliable spike detection, voltage scaling, and pipelining with time-interleaving).

CHAPTER 4

System Design

4.1 Introduction

We now have estimates of all the blocks that are used in constructing a wireless biosignal recording system. In this chapter, we will estimate the minimum achievable system power.

4.1.1 System Candidates

The block diagram of a wireless biosignal telemetry system using digital detection (i.e. detection is post-ADC) in shown in Figure 4.1. A signal from an implanted electrode is measured relative to a reference electrode. The reference electrode is connected at a point which minimizes interference and allows the desired signal to be observed with the best fidelity. The electrodes connect to a recording system that will process the data.

For this analysis, our targeted application is neural spike recording, and one of the first

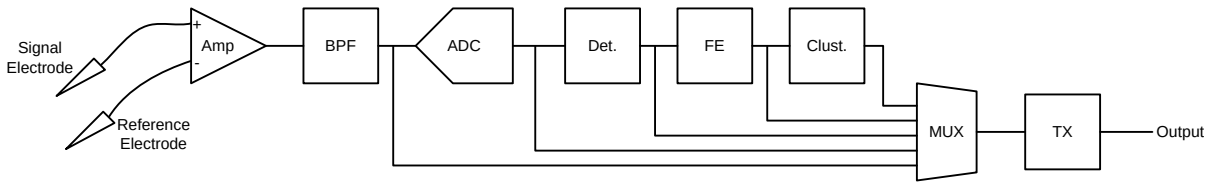

Figure 4.1: Schematic diagram of a wireless biosignal telemetry system, showing options for different output signal modes. Digital detection is shown.

questions is whether analog or digital detection should be used. Spike amplitudes range from implanted electrodes range from 50 μV to 0.5 mV, with background noise in the range 2 μV. This corresponds to an SNR of 13 dB for a minimum-level signal, To ensure the ADC quantization noise does not degrade performance, a margin of 20 dB is added, and the ADC SNR should be greater than 33 dB An automatic gain loop is assumed that will adjust the peak signal levels to stay within the ADC full-scale and a dynamic-range margin of 6 dB is added. The final ADC minimum SNR is then 13+20+6=39 dB, or approximately 7 bits. Since the resolution is low, it is preferable to use digital detection.

The system can be broken in three main sections: (a) the analog front-end (AFE) (b) signal processor (SP) and (c) transmitter (TX). Figure 4.1 shows that all system modes require an analog front-end amplifier and bandpass filter (BPF). The purpose is to boost the signal level and reject unwanted signals. The power and area of the AFE is primarily determined by the noise and bandwidth requirements of the target application, and is essentially independent of the TX and SP design. On the other hand, design of the TX and SP are linked. A high-power complex SP that reduces the datarate allows a reduction in power of the TX. Therefore from a system optimization perspective, we must trade off signal processing complexity with transmitter datarate to minimize the overall power and area. The output of the system can be taken from one of the following: AFE, ADC, detector, feature extractor, or clustering processor. Blocks that are not used can be disabled to save power, or not included in the final system implementation, e.g. if the raw data is taken from the AFE, the ADC and subsequent blocks could be powered down. For a given mode, the corresponding output is routed via a MUX to the TX. Note that the MUX is not required if the hardware is fixed; it is shown here to conceptually illustrate different modes or could be used to provide flexibility on actual hardware. We label the five possible modes as: (a) analog waveform (b) digital raw data (c) spike detection (d) feature extraction and (e) clustering. We will review each case qualitatively with respect to power and area, before calculating the totals. Chapter 2 also described similar systems that have been published,

and we will use these as benchmarks later in this chapter. For the system proposed in this dissertation, we will use estimates presented in Chapter 3 for our optimized blocks.

4.1.1.1 Output Modes

In the first case (a), the analog signal after the BPF can be transmitted directly using an analog transmitter, e.g., a FM transmitter. Since the output signal is an analog modulation it is more sensitive to noise and interference and an inefficient use of bandwidth compared to digital modulation. However when the environment is free of wireless interference and only a small number of channels are transmitted, this is a viable option. The original waveform can be faithfully reproduced up to the fidelity of the amplifier and transmitter. We use this case as the baseline for comparison with other cases.

The next case (b), takes the output of the ADC and transmits the uncompressed raw data. This is similar to case (a), except now the analog waveform is sampled in time and quantized to a level of 8 to 12 bits. The sampling process causes aliasing which can corrupt the desired signal, but can be reduced by the improving the selectivity of the BPF and/or increasing the sample rate. Quantization also introduces additional noise into the signal path and is reduced by increasing the resolution, i.e. the number of bits. Increased selectivity, sample rate, and resolution increases the power and area of the BPF and ADC. However once the signal is digitized, digital error-correction can be applied in the TX to improve the robustness of the transmitted data to unwanted interference. Furthermore, compression algorithms can be applied to the digital data to reduce the datarate required. Lossy compression methods such as ADPCM can be used reduce the data rate to 35% while maintaining a high correlation with the original signal. While it may seem counterproductive to add additional circuitry (i.e. area and power), we claim that the more robust transmitter performance offsets the cost of digitization and signal processing. In this work, we do not use ADPCM, and instead use specialized algorithms.

Both (a) and (b) continuously transmit the signal. However, if the aim is to detect the firing of a neuron instead of recording the complete spike, a binary detector can be used to detect the neuron firing. This is case (c). Detection can capture the samples associated with a spike by creating a recording window around the time that a spike was detected. The "worthless" noise between spikes can then be discarded. This reduces the data rate, and can reduce the overall power of the system if the transmitter does not need to transmit while waiting for a spike detection event. In contrast to the previous two cases, the power now depends on the firing rate of neurons. Spike detection can be performed before or after the ADC. Analog detection, i.e. before the ADC, allows the ADC to be in a low power mode when there is no spike to digitize and save overall power. A spike detector may be triggered by a signal level increase compared to the time-averaged noise level. Since there a time delay between the start of a spike waveform and detection event, the waveform must be buffered so that data before the detection event is not lost. The buffer memory is trivial to implement with digital circuitry in low power and area, but analog memory implementation has a significant overhead. This overhead makes analog detection less attractive depending the difference between (a) the power required for analog detectors with buffering and (b) operating the ADC and digital detector. This was discussed in Section 3.5, where it was concluded that digital detection is preferable when the ADC resolution is less than 10 bits.

Additionally, once the signal has been digitized, complex digital signal processing algorithms can be implemented on chip. These algorithms have been investigated by Gibson and Karkare in [45]. Instead of general-purpose compression algorithms, spike-processing-specific algorithms can be used to achieve higher compression and lower data rates. After spike detection, feature extraction (case d), computes key parameters of a spike waveform such as amplitude and arrival time. However it is not possible to reconstruct the original waveform after feature extraction, but the key parameter are chosen such that sufficient information is computed to discern different spikes, i.e. spikes originating from different neurons. The amount of data reduction can be substantial as a spike may be sampled at 24 kS/s with a

resolution of 10 bits over 3 ms, which results in 720 bits per spike. Feature extraction may reduce this to a small number of parameters requiring a total of 150 bits resulting in 21% of the data required for spike detection. The actual transmitted data rate is dependent on the frequency of spike detection, while in data streaming mode it is independent of firing rate. The additional cost of performing feature extraction on chip must be weighed against the reduced power consumption of the transmitter.

Finally we discuss case (e). Clustering requires additional processing, and is the most complex scheme considered in this dissertation. One way to interpret clustering is that each detected spike waveform is compared to a set of waveforms, and a best match is determined from a relatively small set. The reference waveforms may be continuously updated, or trained by automatic or manual means before data collection begins. Hence clustering typically requires a large amount of memory to store the reference and detected waveforms. The data required for each classified waveform can be as low as 6 bits which results in a very low data rate, and would allow recording from thousands of channels while keeping the transmitter data rate within practical limits.

4.1.1.2 Transmitter Operation

The maximum allowable datarate for the transmitter is limited by the bandwidth, power, and bit error rate. In an effort to make transceiver design practical, 10 Mbps datarate has been chosen, based on the fact that bit error rate (BER) becomes unacceptable beyond 10 Mbps. This 10 Mbps limit can be overcome with more complex transmitters and receivers, but is considered beyond the scope of this work [80].

The selected signal is routed to the transmitter. In the case of analog waveform mode, the transmitter constantly transmits data. If the waveform is digitized, the data can be buffered, which then allows the transmitter to be enabled only when needed. This can be employed in cases (b)-(e). As previously discussed, a maximum transmitter data rate of 10 Mbps is used.

The recorded neural signals are buffered, and a packet of data transmitted every T_{LATENCY}. It is assumed that the power amplifier is able to be enabled instantaneously, and hence its power is proportional to datarate. The synthesizer, on the other hand, has a start-up time, as well as time operating when data is being transmitted. The transmitter start-up time is given by T_{START}. Hence there is time before each packet transmission that the synthesizer is wasting power, i.e. not transmitting bits. It is assumed the latency is much larger than the startup time. In each period T_{LATENCY}, the system will buffer an average of N_{BUFFER} bits of data. The amount of data depends on the selected mode. The time T_{SYNTH} that the synthesizer is on is given by $T_{\text{START}}+N_{\text{BUFFER}}/(10\text{ Mbps})$. The duty cycle of the synthesizer is $T_{\text{SYNTH}}/T_{\text{LATENCY}}$, from which we can determine average power. It should be noted that when the data buffer fills rapidly enough the synthesizer will remain on as there will not be enough time to restart it before the next packet. Each recording channel generates 192 kbps for raw data, and can be reduced depending on the mode as shown in Table 4.1, which in turn reduces the duty cycle of transmitter operation.

Table 4.1: Data rates for different modes.

Mode	Raw Data	ADPCM	DET	FE	CLUST
kbps	192	64	48	10	0.4
Power	0	1	1	4	8

From this discussion, it is clear that using compression reduces the power required for the transmitter and that increasing the latency reduces overhead of the synthesizer startup time. Since ADPCM uses similar power to DET, but has a higher data rate, we will exclude ADPCM from further analyses.

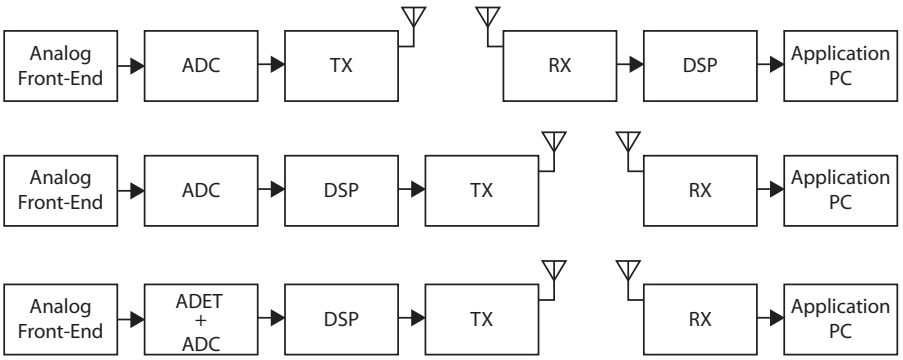

Figure 4.2: Different system configurations for wireless neural recording: (a) raw data (b) spike detection (c) digital signal processing.

4.1.1.3 System Summary

Published systems were summarized in Chapter 2. Based on these systems which had from 4 to 128 channels, we will estimate their power and area as the number of channels increases up to a maximum channel count of 1000. In the next section, we will examine these systems in more detail. We will then construct our system using the blocks described in Chapter 3, and compare the proposed system to the existing work as a function of channel count.

4.2 System Power Estimates

Now that we have power estimates for the main blocks of the wireless telemetry system, we can calculate the power dissipation for each system mode. Three different DSP methods are compared for the system.

The next step is to determine whether it is beneficial to use compression while taking power and digital-system complexity into account. To do this, we will calculate the power for each option to determine which has the lowest power.

The system power is determined by summing the contributions from the AFE, one of the

DSP implementations, and TX. The power for the AFE is based on noise specifications and is constant for all modes. The power dissipation and data compression of the DSP increases across the detection, feature extraction and clustering modes. The improved compression allows a reduction in TX data rate. The block power is summarized in Table 4.2.

Table 4.2: Power for each block.

Block	Power	Unit	Note
AFE	6	μW/chan	20 kHz, 3 μV noise
DSP Det.	1	μW/chan	
DSP FE	4	μW/chan	
DSP Clust	8	μW/chan	
Synth	3	mW	500 μs startup. TX blocks
UPX	1.5	mW	are shared over all channels.

To calculate the power for a system with a given number of channels, we add the per-channel entries for the AFE and DSP multiplied by the number of active channels. The aggregate data rate from all channels can be calculated based on the average data rate per channel. With this data rate, the duty cycle and power dissipation of the transmitter can be calculated. A latency of 1 ms is used for these calculations. The total system power is then the sum of these components.

The first case that we examine is raw data mode (Fig. 4.2a). The calculations are shown graphically in Figure 4.3. In this mode, we can transmit up to 50 channels before hitting the 10 Mbps limit. For a small number of channels, the power is slightly lower than 2 mW. The synthesizer is enabled 500 μs before the transmitter. When 25 channels are active the transmit time is also approximately 500 μs, i.e. the complete transmit window of 1 ms is used. For more than 25 channels, it is necessary for the synthesizer to be on constantly, and as such, the average power of the synthesizer does not increase as more channels are

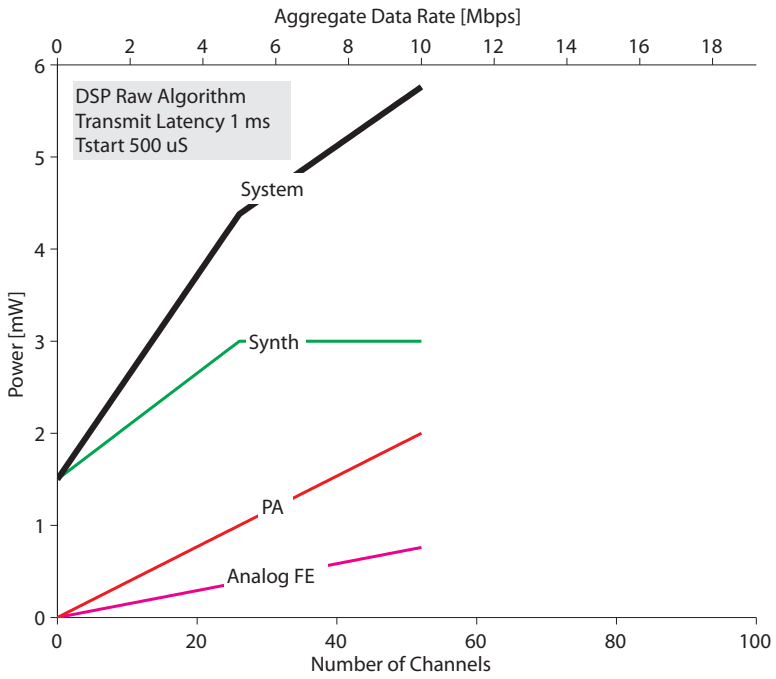

Figure 4.3: Raw streaming.

added. This is the reason for the reduced slope with more than 25 channels. With 192 kpbs per channel, a 10 Mbps transmitter can support approximately 50 channels of uncompressed data. For less than 2.5 Mbps, the synthesizer power, due to its slow start-up, is significant compared to the UPX power. This can be reduced by using a faster-start-up synthesizer or a longer buffer (leading to longer latency). The total power for 52 channels is 5.8 mW, or 110 μW per channel.

Introducing spike detection and only transmitting spike data (Fig. 4.2b) reduces the amount of transmit data from 192 kbps to 48 kbps. Figure 4.4 shows that the DSP power required is negligible and the total power is reduced to 54%. Since the detection DSP power is low, the lower data rate provides significant power saving in the transmitter. With this approach, 100 channels is now possible using a 4.8-Mbps link. In this analysis a 100 Hz spike detection rate was assumed. Lower spike rates would provide even greater power savings as only detected spikes are transmitted.

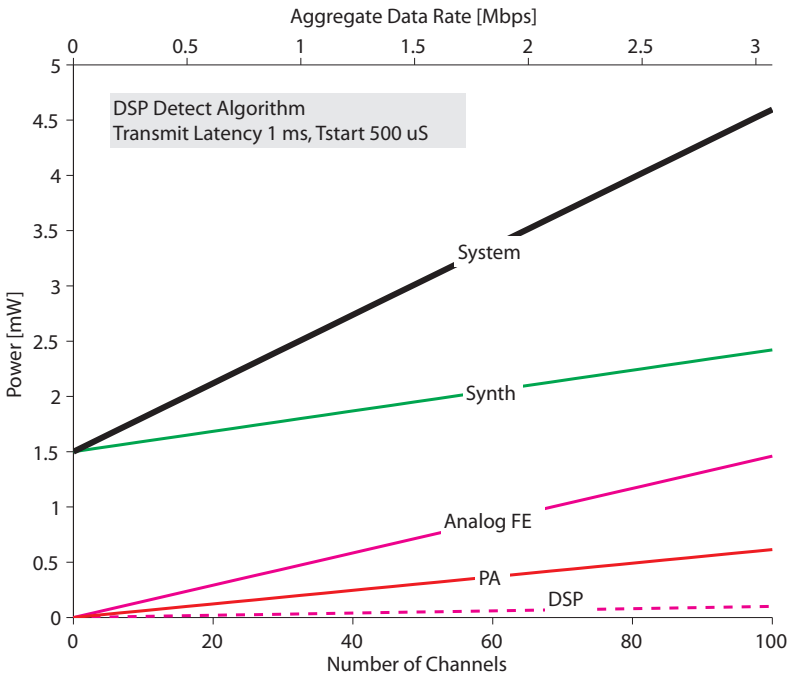

Figure 4.4: DSP-detection mode.

Similarly, using feature extraction and clustering reduce the overall system power compared to raw data mode. In the feature extraction case, both the DSP and TX power are low which indicates that feature extraction is an excellent choice for integration. While clustering can further reduce the data rate, the DSP power now approaches the power of the AFE. Hence it is recommended that clustering should only be used in situations that require very high number of channels. An exception would be devices that need the identification of single neuron activity on the implanted device; future brain-machine interfaces may require such functionality.

With the estimates given so far, it is clear that the synthesizer is a bottleneck in achieving even lower power. If this startup time of the synthesizer can be eliminated, the power would follow a curve similar to the power amplifier, i.e. pass through the origin of our plots. Two approaches are to develop a synthesizer that can "hold" its state while going into a low-power state and return quickly to stable operation, or use an oscillator in open-loop mode, with

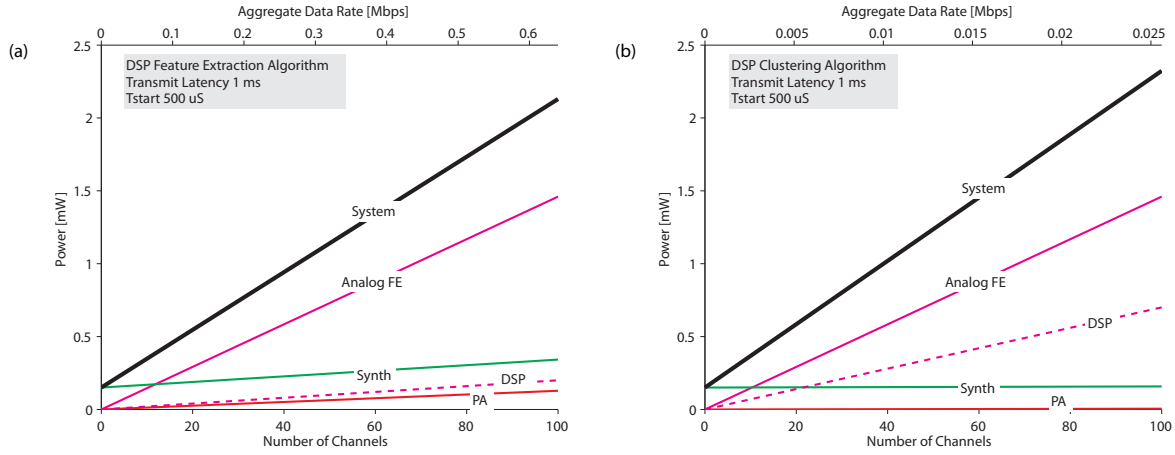

Figure 4.5: Power dissipation of feature extraction and clustering modes.

digital control to set the frequency. For simplicity, the second option is considered here. There are three issues that must be considered (a) an open-loop oscillator has high close-in phase noise, and this must not overwhelm the signal that we transmit (b) the startup time must be much shorter than we currently have and (c) the receiver be able to track the frequency drift due to the open-loop operation. Figure 4.6 shows the performance an LC oscillator that dissipates 1 mW from a 1.2 V supply. The measured phase noise of the oscillator yields a 42 dB SNR, which indicates that phase noise will not limit performance for a BPSK signal that requires 6 dB SNR for demodulation. The startup time, due to thermal noise in this simulation, shows that the output is stable in approximately 40 ns after the bias current is applied. The receiver design to track the open-loop oscillator is not considered here. The use of such an oscillator could reduce the power of the overall system substantially.

Another question that arises is can high (i.e. 60 GHz) frequency transmitters be used to transmit the full raw-data stream and avoid the need for compression. Leeson's equation (Eq. 4.1) indicates that with the same modulation data occupying frequency offsets up to ~10 MHz (centered around the carrier) the power of the oscillator must be increased. This is because

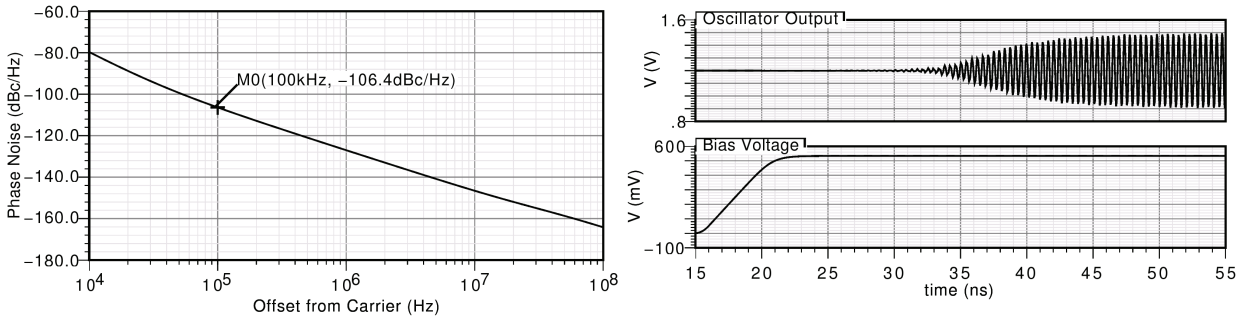

Figure 4.6: Performance of open-loop LC oscillator.

if the carrier frequency f_0 is increased, the power P_{avs} must be increased to compensate. For this reason, it is preferable to keep the transmitter frequency low. However, other constraints such as the presence of interference also play a role in the selection of frequency.

$$\mathcal{L}(f_{\text{m}}) = \frac{F \cdot k \cdot T}{2 \cdot P_{\text{avs}}} \left[1 + \left(\frac{f_0}{2 \cdot f_{\text{m}} \cdot Q_{\text{L}}} \right)^2 \right] \tag{4.1}$$

where $\text{L}(f_{\text{m}})$ is the phase noise in dBc/Hz, F is empirical constant for the oscillator topology, P_{avs} is the average power through the resonator, f_0 is the carrier frequency, f_{m} is the frequency offset from the carrier, and Q_{L} is the quality factor of the resonator.

The final question is can we increase the latency of the transmitter, so that the synthesizer could remain off for longer durations. Longer buffering requires more memory, and we can estimate how much memory we could add before it becomes 10% of the analog frontend power. The leakage current of memory in 65-nm CMOS is approximately 104 pA. With 6 μA in the analog frontend, 104 pA leakage current in a standard-cell memory block, and 8 bits per sample, ~7200 samples can be stored within a reasonable power limit. At 24 kS/s, this would provide a buffer depth of 300 ms. This could reduce the synthesizer power significantly compared to a latency of 1 ms. With a 2.6 μm^2 standard-cell memory layout, the area of memory to hold 7200 samples would be 0.15 mm^2. Unfortunately, this is similar to the area of the analog frontend, and it may be prohibitive to use this technique.

4.3 Conclusion

We have shown a system design for wireless telemetry that enables a large number of channels. Previously published work is shown in Fig. 4.7. Analog implementations [33, 83] are suitable for low channel counts, and have higher power compared to the other systems. Spike-detection systems [35, 43] show high channel counts and/or lower power per channel. Finally, spike sorting [41] demonstrates the potential (and even necessity) for increasing local-digital-signal processing to facilitate a low-power system. The estimates for different modes of our proposed system are also shown in the figure. Four cases (a) raw digital data, (b) spike detection, (c) feature extraction, and (d) clustering as shown. All realizations yield lower power than the corresponding published work.

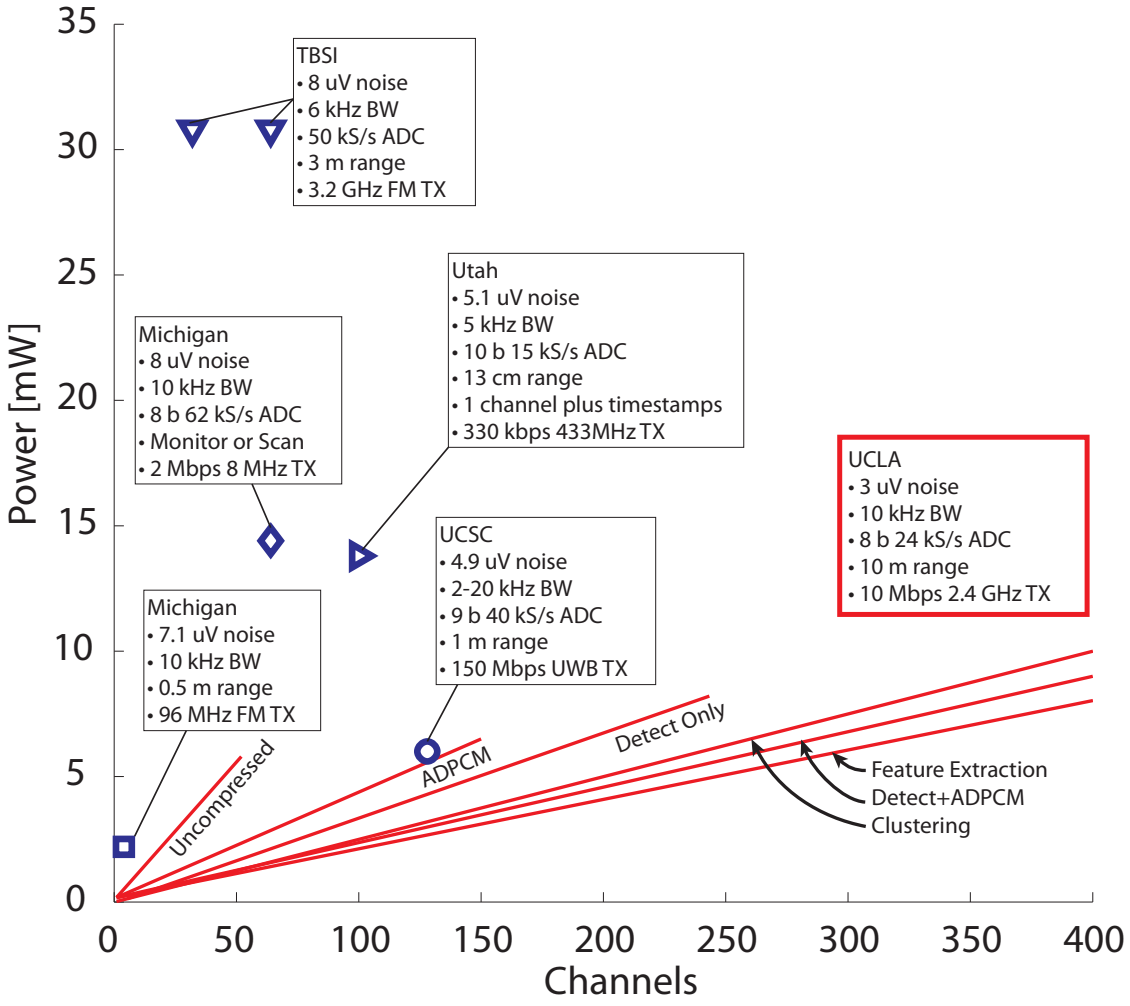

Figure 4.7: Implemented Wireless Neural Recording Systems: Michigan [33, 43], Utah [35], UCSC [41], TBSI [42]. Designs are not normalized (for bandwidth, input referred noise, range, and features.)

CHAPTER 5

Implementation

5.1 Introduction

Using the design methodology described in Section 3.3.8, the schematic complete amplifier is shown in Figure 5.1. The design is implemented in TSMC's 65-nm CMOS process with mixed-signal device process options (MIM capacitors and unsalicided polysilicon resistors). Stage 1 is the largest amplifier since it requires the lowest noise. Stage 2 and 3 use the same amplifier core, but stage 2 does not use ac coupling to maximize the gain. Stage 3 is ac coupled to prevent the dc offset saturating the amplifier. Stage 4 is included for testing, and operates from a higher supply to allow a source follower output to drive large off-chip load capacitance. The output of each stage is also connected to an output pad, so that debugging could be done if necessary, as well as adjusting the frequency response with external components. Biasing is not shown. For testability, all bias currents are externally generated. This allows debugging, as well as the ability to use external sources to adjust the bias and investigate the noise-power trade-off of the amplifier.

The schematics of the amplifiers are shown in Fig. 5.2 and Fig. 5.3. Current source loads were added instead of pure resistive loads to allow a higher gain. Unfortunately, the noise of the current sources was not measured before tapeout and adds a significant amount of noise, but the noise was still close to the original target of 2.5 μV.

Simulation shows that the total noise for the amplifier is 3.2 μV integrated over a band from 100 Hz to 10 kHz with a supply current of approximately 8 μA. The ac voltage gain

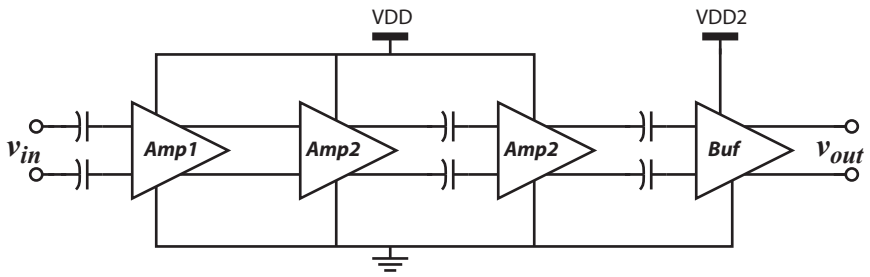

Figure 5.1: Architecture of cascaded amplifiers

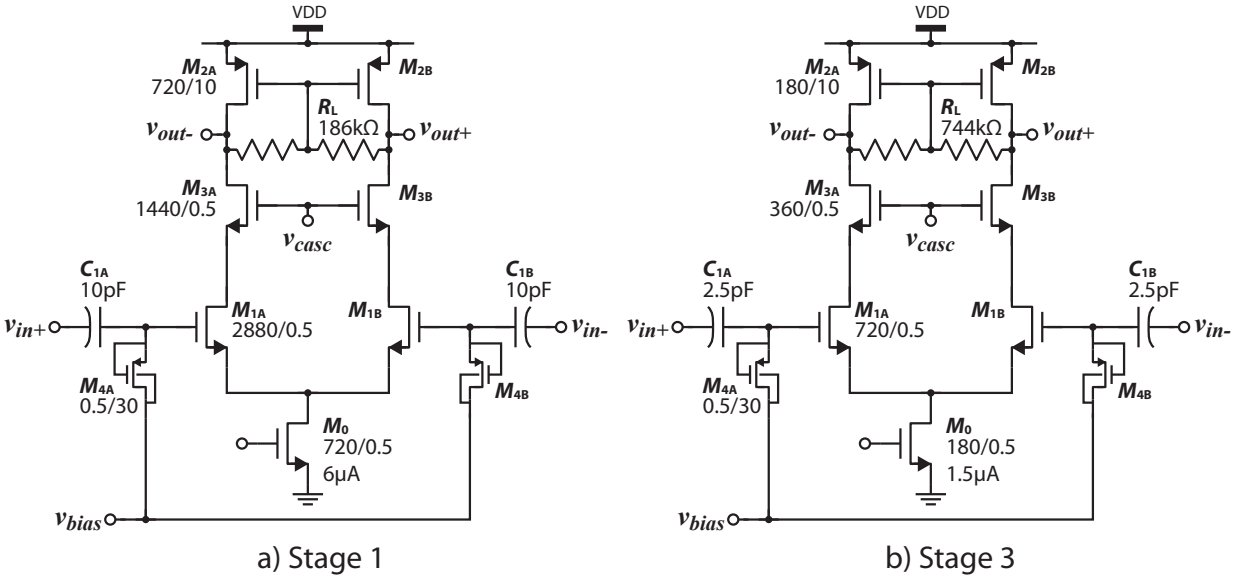

Figure 5.2: Stages 1–3 of the designed amplifier.

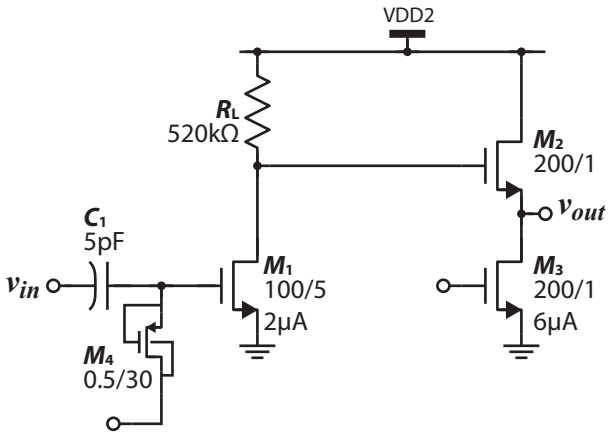

Figure 5.3: Stages 4 of the designed amplifier.

is 58 dB to the output of the 3rd stage, and 78 dB after the buffer. Simulated distortion is less than -51 dBc.

The layout of the amplifier is shown in Fig. 5.4.

5.1.1 Test PCB

A schematic diagram of the printed circuit board used for testing is shown in Figure 5.5. A low-frequency source is used to generate a sinusoidal input. A resistive attenuator is used to reduce the amplitude to the micro-volt level required for the amplifier. The amplifier is driven signal ended at the positive input terminal with the other terminal grounded, which is similar to how it would be used in practice. A commercial differential amplifier is used to convert the output to a single-ended signal before driving the spectrum analyzer. Since the spectrum analyzer has a 50Ω input impedance, a resistive divider is also used at the output of the SE-to-differential stage to keep the load resistance above the minimum required.

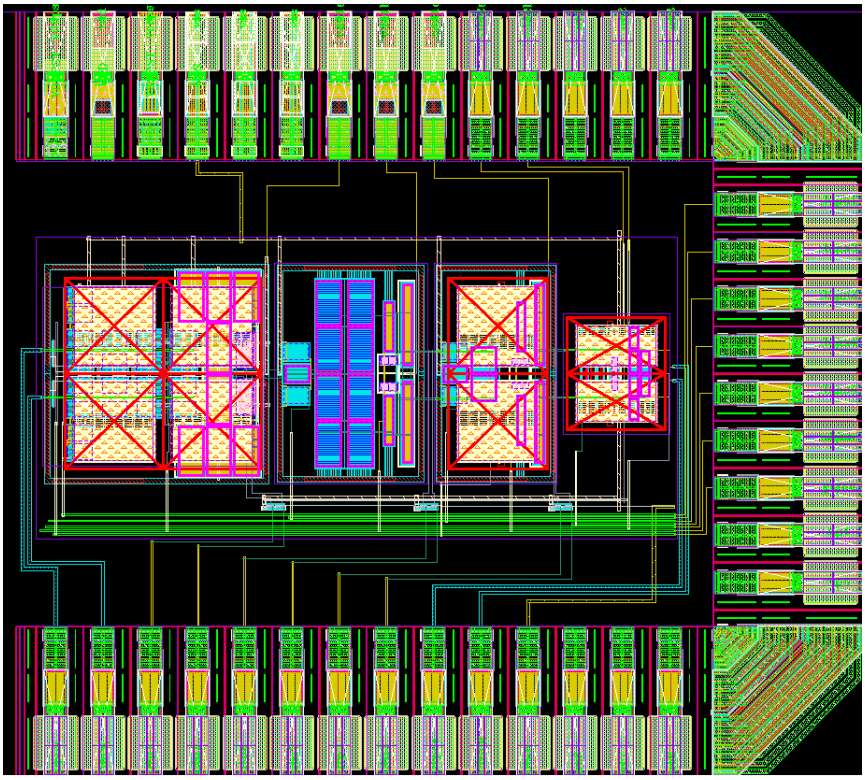

Figure 5.4: Layout of a cascaded open-loop amplifier topology. Total area is 0.2 mm^2.

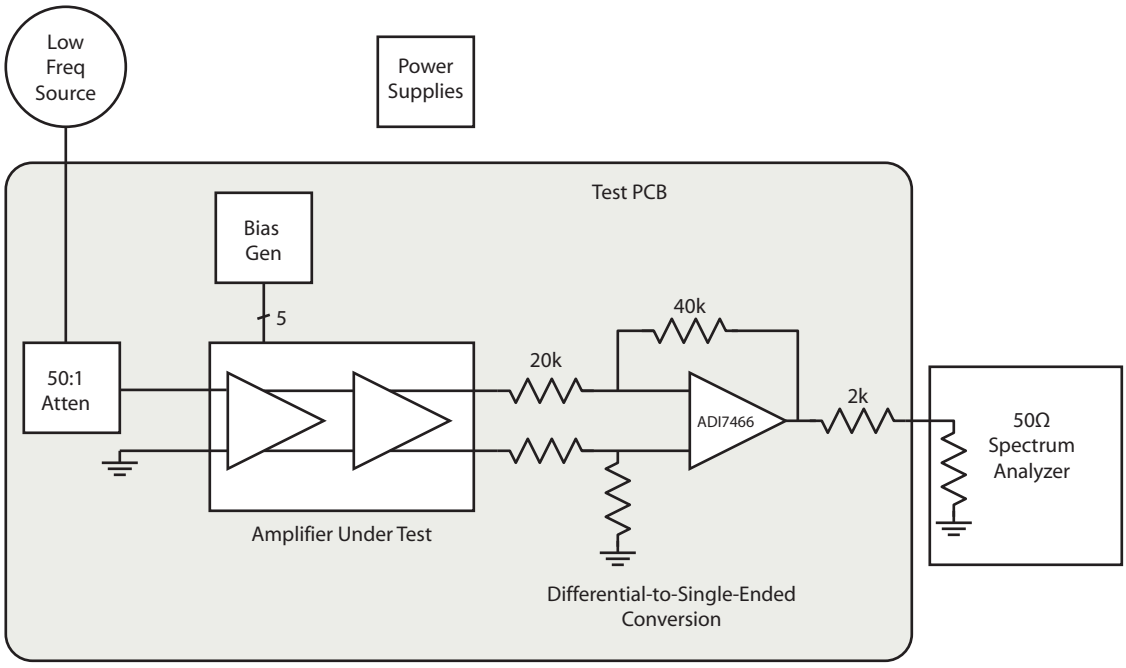

Figure 5.5: Schematic diagram of the printed circuit board and test equipment.

5.2 Measurement Results

The frequency response and gain for the amplifier are shown in Fig. 5.7. The gain is approximately 5 dB lower than simulated. It is unclear why the gain is lower, but gain depends on bias current, but measuring the total supply current accurately was difficult. A parasitic gate leakage current in the ESD diodes of approximately 5.5 μA was estimated from simulation (Fig 5.6). Since the gain is on the order of $(g_{\mathrm{m}} \cdot R_{\mathrm{L}})^4$, an error of 15% in the 17 μA supply current measurement could account for this gain discrepancy, and could be affected by the ESD diode leakage. However, the absolute gain of the circuit was not a primary concern. The noise however matches reasonably well with expectation. Noise was integrated up to 10 kHz, with the assumption noise higher than this frequency would be removed by filtering in the DSP.

The noise versus supply current is shown in Fig. 5.8. The bias currents in the first three

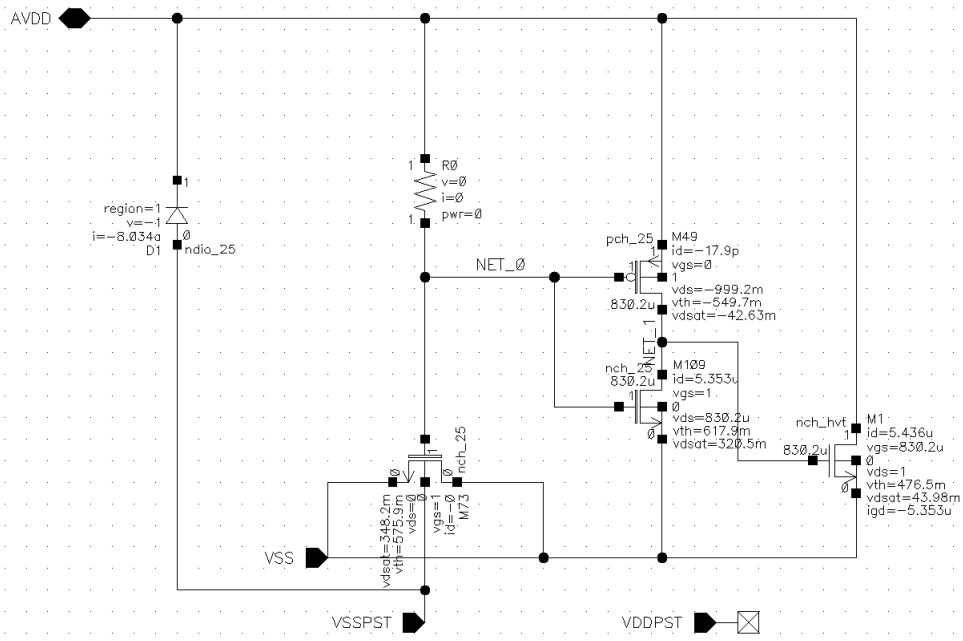

Figure 5.6: Supply leakage current through M1 of approximately 5.5 μA.

stages were adjustable. Reducing the bias currents causes the voltage gain to reduce and the noise increase. The output stage bias was kept constant. It is expected that the noise should be proportional to the square of the supply current. However, with a constant current in the output stages, there is a minimum power P_{min} that the amplifier dissipates. As shown in Fig 3.5, the amplifier noise that can be tolerated, based on SNR considerations, varies with the signal conditions. With a weak spike signal, the amplifier would operate at minimum noise, and consequently, highest supply current of 17 μA. With a stronger signal, the allowable noise is also higher, which corresponds to lower power. For the prototype amplifier, Fig. 5.8 shows the noise-power tradeoff.

In a large array of amplifiers, with a range of spike amplitudes, there can be power savings overall if each amplifier current is set optimally. For instance, if all electrodes received a weak signal, then all amplifiers would be required to operate at maximum current. On the other hand, if a range of amplitudes were present, e.g. having a Gaussian distribution with a mean and standard deviation of 400 μV and 100 μV respectively, the overall power would be 62%

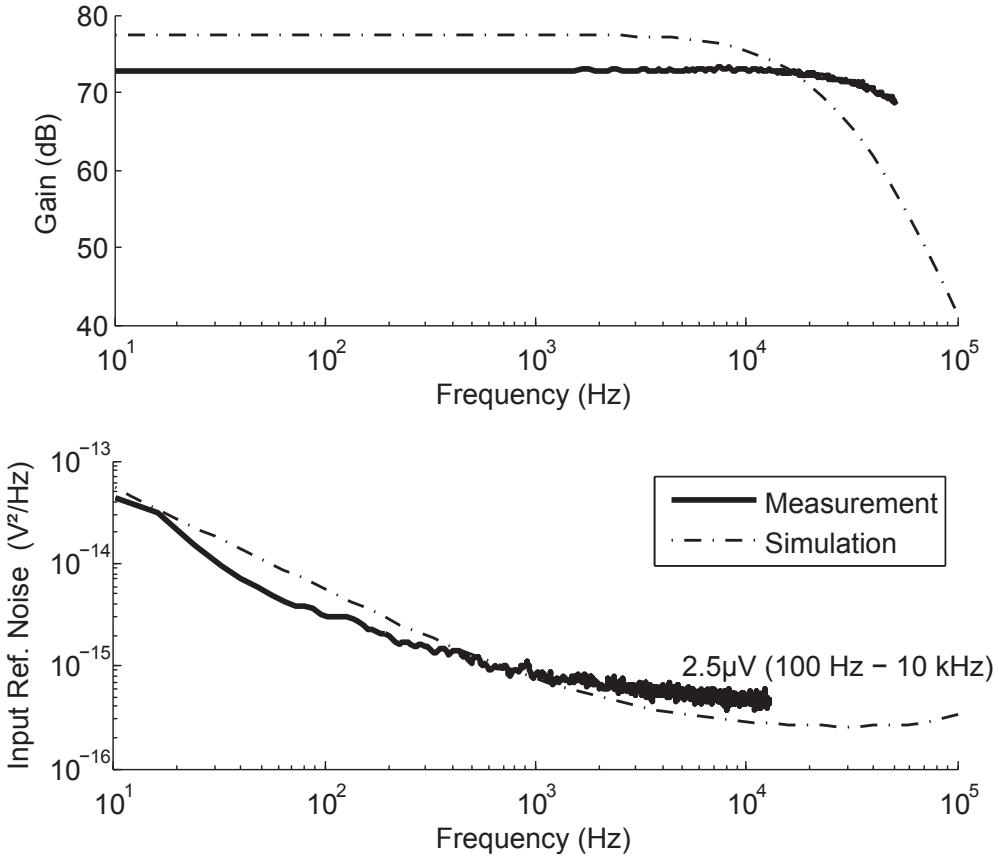

Figure 5.7: Measured gain and noise compared to simulation, for 17 μA supply current.

lower. This is illustrated in Fig. 5.9.

The third harmonic distortion is better than -35 dBc when measured at the spectrum analyzer, although simulation shows this is limited by the output buffer. Without the output buffer, distortion is better than -41 dBc (Fig. 5.10).

5.3 Summary

The performance of the amplifier is compared in Tab. 5.1. This chapter described the implementation of a low-power amplifier for neural spike recording. It was implemented in a 65 nm CMOS process with a 1 V supply. The supply current is adjustable from 2 to 17 μA,

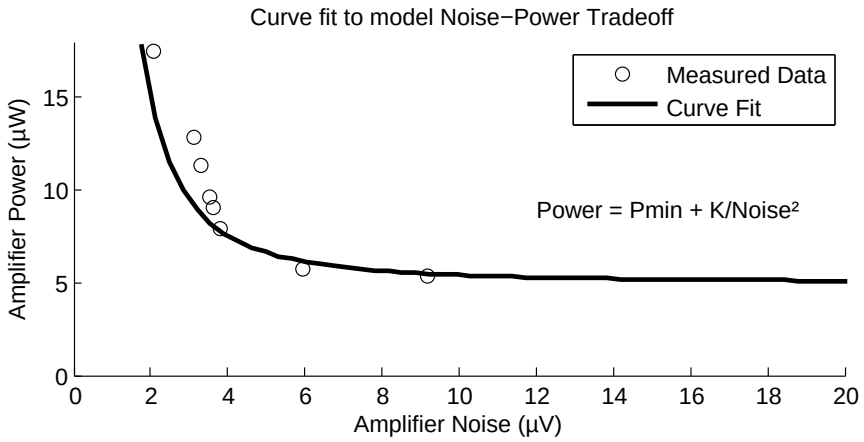

Figure 5.8: Noise versus amplifier power.

with corresponding input referred noise of 2 to 10 μV, and gain from 24 to 73 dB. The active area is 0.17 mm^2. The adjustable bias current would allow power dissipation to scale with received signal strength, which could potentially save power over a large array of amplifiers. Sharing the bias generation between several blocks could reduce the power dissipation as well.

Finally, the amplifier is compared to other recent work in Table . Since NEF does not account for supply voltage, the comparison uses *Normalized Power* to compare power efficiency. Normalized Power estimates the power required for each amplifier, and scaled for different bandwidth and input referred noise, and is calculated by

$$Normalized\ Power = (9.5\ \text{kHz}/Bandwidth) \cdot (Noise/2.5\ \mu\text{V})^2$$

. This work was fabricated in a process that allows dense integration of digital circuitry, but also has high flicker noise, which leads to large device area for the amplifier. Both open-loop topologies show similar power efficiency, but the best efficiency is observed for an inverter-based topology. The implemented amplifier also has higher gain which is implemented by

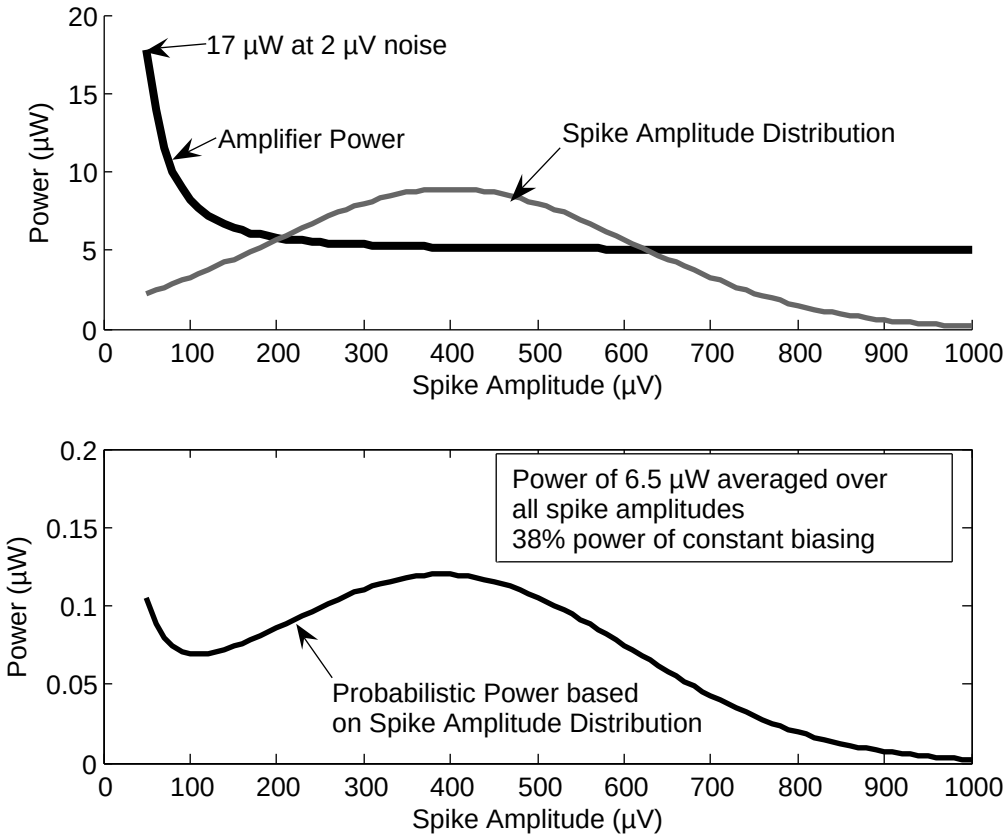

Figure 5.9: Estimated power saving for optimal bias currents based on spike amplitude.

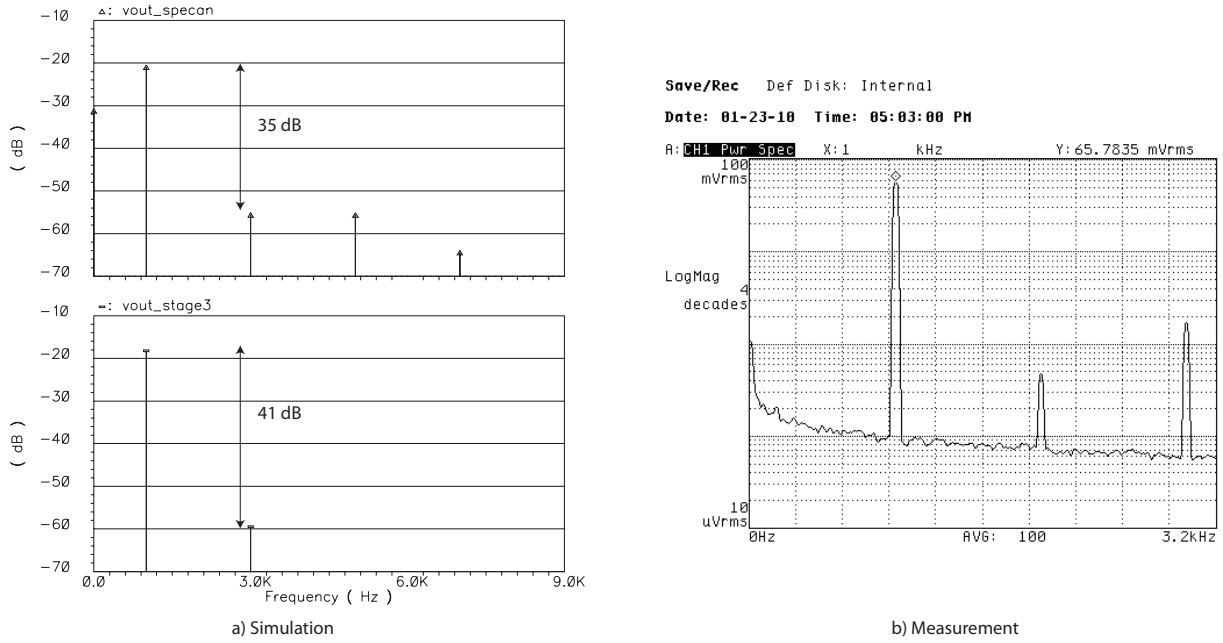

Figure 5.10: Third harmonic distortion comparison between simulation and measurement.

additional gain stages, which increase the noise and power.

Author	Wattanapanitch	Chang	Rai	This Work	
Ref	[59]	[84]	[85]		
Noise	3.1	1.7	1.8	2.5	μV
Area	0.16	0.043	0.062	0.17	mm^2
Power	7.60	1.00	12.5	17.2	μW
Bandwidth	5.3	0.292	11.5	9.5	kHz
VDD	2.8	2.5	1.0	1.0	V
Technology	0.35 μm	0.25 μm	0.13 μm	65 nm	
Topology	OTA	Open-Loop	Inverter	Open-Loop	
Voltage Gain	40.0	40.0	38.3	74	dB
Normalized Power	20.9	15.0	5.35	17.2	μW
Noise PSD	42	85	18	21	nV/rtHz

Table 5.1: Performance comparison with other amplifiers.

CHAPTER 6

Conclusions

Circuitry similar to that described in this work is necessary to address significant research problems in neuroscience, ranging from understanding neural processes to repair brain function, to designing brain-machine interfaces to aid amputees and stroke victims. Before such major issues can be solved, it is necessary to have a low power system which will not impair the health or mobility of human subjects, and allow long term measurements to be taken. High density, state-of-the-art micromachined electrodes allow observations from a large number of electrode implant sites, and this work elevates the efficiency of associated electronics to match this capability.

State-of-the-art systems have significant limitations, either in terms of area or power dissipation, which subsequently limit the total number of channels supported, and thus limited their effectiveness for neuroscience research. To find the lowest power system several system configurations were considered, and the building blocks for these systems were analyzed. With advances in CMOS technology and low-power design techniques, the improved performance of the digital sections of the system have been previously demonstrated. The analog front-end and transmitter were the focus of this work since these areas have yet to match the performance of the other blocks. Furthermore, the reduced data rate provided by the DSP can be leveraged to use a lower power transmitter. The noise of the analog amplifiers was minimized for a given power budget, and a methodology was described for the amplifier design, which highlights the key trade-offs in terms of area, power and noise, that can be used for optimization by the system designer. A prototype 1 V amplifier was also fabricated to

measure performance in a 65-nm CMOS process that avoided the high-supply voltage commonly used in previous designs. The low-supply-voltage operation allows integration with low-power digital signal processing on a compact single-chip design. The amplifier noise-efficiency performance was comparable to other recently published amplifiers. The range of signal and noise conditions over a large number of channels can be incorporated into the system-level optimization by setting each amplifier current to meet the required SNR while using minimum power. The programmable bias current feature was also included in the prototype.

Finally, with power estimates of all the blocks in the system, the different system configurations were analyzed to determine whether raw data, spike detection, feature extraction or clustering was the most power efficient. This study focused on neural spikes, and found that a system that uses feature extraction yields the lowest overall power, can support 400 channels with a practical wireless link, while consuming approximately 8 mW. This is an improvement on existing work, while also providing flexibility for different modes, and provides significant capability for future neuroscience research.

APPENDIX A

A Short Summary of EKV Model

A.1 I_S Definition

The characterization centers on the following normalization:

$$I_S = \frac{2}{\kappa} \cdot \mu \cdot C_{ox} \cdot V_T^2 \cdot \frac{W}{L} \tag{A.1}$$

Next the Inversion Coefficient IC, is defined as:

$$IC = \frac{I_D}{I_S} \tag{A.2}$$

When IC is equal to 1, the device is in moderate inversion. Strong and Weak inversion are approximately occur when IC is >10 and <0.1 respectively. It is also common to characterize inversion level by the current density I_ρ. Inversion level will be described as IC or current density, as the two terms are convey the same concept, but have different mathematical definitions.

$$I_\rho = I_D/W = \frac{2 \cdot \mu \cdot C_{ox} \cdot V_T^2}{\kappa \cdot L} \cdot IC \tag{A.3}$$

A.2 Transconductance

The transconductance of the device can be approximated by:

$$\begin{aligned} g_{\mathrm{m}} &= \frac{2 \cdot \kappa \cdot I_{\mathrm{D}}}{V_{\mathrm{T}}} \cdot \frac{1}{1 + \sqrt{1 + 4 \cdot IC}} \\ &= 4 \cdot \mu \cdot C_{\mathrm{ox}} \cdot V_{\mathrm{T}} \cdot \frac{W}{L} \cdot \frac{IC}{1 + \sqrt{1 + 4 \cdot IC}} \end{aligned} \tag{A.4}$$

The transconductance approaches its asymtotic values given by equations A.5 and A.6. Note that A.6 overestimates transconductance in strong inversion ($2 \cdot \kappa$ is greater than unity), but in practice such high levels of IC are not achievable.

$$g_{\mathrm{m}} = \frac{\kappa \cdot I_{\mathrm{D}}}{V_{\mathrm{T}}} \tag{A.5}$$

$$\begin{aligned} g_{\mathrm{m}} &= \sqrt{2 \cdot \kappa} \cdot \sqrt{2 \cdot \mu \cdot C_{\mathrm{ox}} \cdot \frac{W}{L} \cdot I_{\mathrm{D}}} \\ &= \sqrt{2 \cdot \kappa} \cdot g_{\mathrm{m,square\ law}} \end{aligned} \tag{A.6}$$

A.3 Capacitance

The capacitance of a MOSFET is given by:

$$C_{\mathrm{g,strong}} = C_{\mathrm{gs,strong}} = \frac{2}{3} \cdot C_{\mathrm{ox}} \cdot W \cdot L \tag{A.7}$$

$$C_{\mathrm{g,weak}} = C_{\mathrm{gs,weak}} + C_{\mathrm{gb,weak}} = (IC + 1 - \kappa) \cdot C_{\mathrm{ox}} \cdot W \cdot L \tag{A.8}$$

Shown in (Figure A.1), a series combination of $C_{\mathrm{g,strong}}$ and $C_{\mathrm{g,weak}}$ can be used to approximate the capacitance from weak to strong inversion. [86]

Using equations A.7, A.8, and A.2, the input capacitance C_{in} is:

$$\begin{aligned} C_{\text{in}} &= C_{\text{g,strong}} \| C_{\text{g,weak}} \\ &= \frac{\kappa \cdot I_{\text{D}} \cdot L^2}{\mu \cdot V_{\text{T}}^2} \cdot \frac{IC + 1 - \kappa}{IC \cdot (3 \cdot IC - 3\kappa + 5)} \end{aligned} \tag{A.9}$$

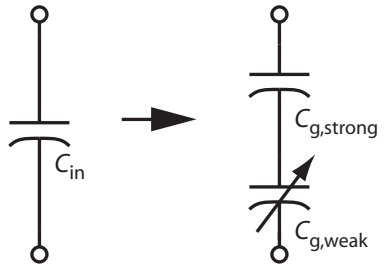

Figure A.1: Interpolation of Strong and Weak Inversion capacitances.

Additionally, the gate-to-source overlap capacitance can be included as C_{ov}:

$$C_{\text{ov}} = C_{\text{gs,overlap}} \cdot W = \frac{C_{\text{gs,overlap}}}{C_{\text{ox}}} \cdot \frac{\kappa \cdot L \cdot I_{\text{d}}}{\mu \cdot V_{\text{T}}^2 \cdot IC} \tag{A.10}$$

A.4 Saturation Voltage $V_{\text{DS,sat}}$

The FET is considered to be in saturation mode when $V_{\text{ds}} > V_{\text{ds,sat}}$ [87], where

$$V_{\text{ds,sat}} = 2 \cdot V_{\text{T}} \cdot \sqrt{IC + 0.25} + 3 \cdot V_{\text{T}}. \tag{A.11}$$

With this definition, g_{m} can be expressed as

$$g_{\text{m}} = \frac{2 \cdot I_{\text{D}}}{n \cdot (V_{\text{ds,sat}} - 2 \cdot V_{\text{T}})}. \tag{A.12}$$

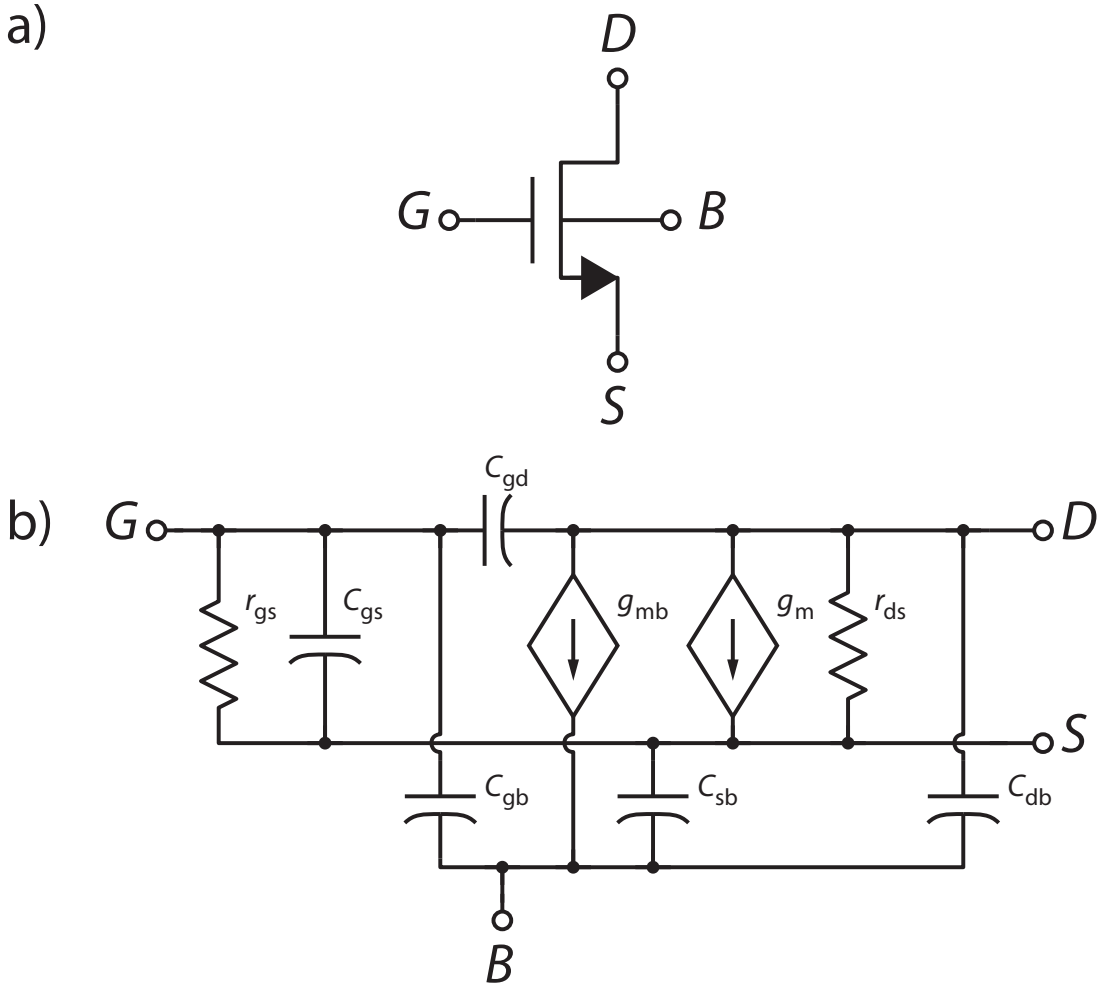

Figure A.2: Small-Signal model of a MOSFET.

A.5 Modelling of Transistors versus IC

Most modelling of MOS transistors focuses on strong inversion. However, low-power applications often benefit from subthreshold operation. The following section briefly describes the EKV model [86] that approximates device behavior from weak inversion to strong inversion. This model will be used to search for an optimal operation point.

Symbol	Description	Value	Units
k	Boltzmann Constant	1.380×10^{-23}	J/K
T	Temperature	310	K
C_{ox}	Oxide Capacitance (350nm CMOS)	4.5	fF/μm^2
μ	Charge Carrier Mobility	–	μm^2/V/s
μ_{N}	Mobility in N-type Material	350×10^8	μm^2/V/s
μ_{P}	Mobility in P-type Material	100×10^8	μm^2/V/s
γ	Channel Noise Coefficient	1/2 to 2	–
κ	Subthreshold Swing	0.7	–
n	Subthreshold Slope (=$1/\kappa$)	1.4	–
V_{T}	Thermal Voltage @ 37°C (= $k \cdot T/q$)	26.8	mV
q	Electron Charge	1.609×10^{-19}	C
W	Transistor Width	–	μm
L	Transistor Length	–	μm
I_{S}	Normalization Current	1.104 $\times W/L$ (NMOS)	μA
I_{D}	Transistor Drain-to-Source Current	–	μA
IC	Inversion Coefficient	≈0.1–10	–
I_ρ	Current Density (= I_{D}/W)	≈1–100	μA/μm
K_{CAP}	Capacitor Density	1	fF/μm^2
K_{FET}	(Input Device Area)/(Amplifier Area)	1/3	–

Table A.1: Constants and Variables used to characterize Amplifiers.

References

[1] S. D. Shorvon, "The epidemiology and treatment of chronic and refractory epilepsy," *Epilepsia*, vol. 37, no. 2, pp. 51–53, 1996.

[2] J. R. Wolpaw, N. Birbaumemr, D. J. McFarland, G. Pfurtscheller, and T. M. Vaugan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, pp. 767–791, 2002.

[3] P. J. Rousche, K. J. Otto, M. P. Reilly, and D. R. Kipke, "Single electrode microstimulation of rat auditory cortex: an evaluation of behavioral performance," *Hearing Research*, vol. 179, no. 1-2, pp. 62–71, 2003.

[4] J. K. Chapin, K. A. Moxon, R. S. Markowitz, and M. A. L. Nicolelis, "Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex," *Nature Neuroscience*, vol. 2, no. 7, pp. 664–670, 1999.

[5] G. Fridman, H. Blair, A. Blaisdell, and J. W. Judy, "Somatosensory feedback for brain-machine interfaces: Perceptual model and experiments in rat whisker somatosensory cortex," in *Proceedings of the 3rd International Conference IEEE Engineering in Medicine and Biology Society Conference on Neural Engineering*, Kohala Coast, HI, USA, May 2007, pp. 2–5.

[6] S. W. Badelt, A. P. Blaisdell, H. T. Blair, and J. W. Judy, "Intention and animal models for brain computer interfaces," in *BMES Annual Fall Meeting*, Los Angeles, CA, USA, September 2007, pp. 26–29.

[7] M. A. Nicolelis, "Actions from thoughts," *Nature*, vol. 409, no. 6818, pp. 403–7, 2001.

[8] D. M. Taylor, S. I. Helms-Tillery, and A. B. Schwartz, "Direct cortical control of 3D neuroprosthetic devices," *Science*, vol. 296, no. 5574, pp. 1829–1832, 2002.

[9] J. Wessberg, C. R. Stambaugh, J. D. Kralik, P. D. Beck, M. Laubach, J. K. Chapin, J. Kim, S. J. Biggs, M. A. Srinivasan, and M. A. L. Nicolelis, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," *Nature*, vol. 408, no. 6810, pp. 361–365, 2000.

[10] G. Santhanam, M. D. Linderman, V. Gilja, A. Afshar, S. I. Ryu, T. H. Meng, and K. V. Shenoy, "HermesB: A continuous neural recording system for freely behaving primates," *IEEE Transactions in Biomedical Engineering*, vol. 54, pp. 2037–2050, 2007.

[11] D. Liu, J. Diorio, J. C. Day, D. D. Francis, and M. J. Meaney, "Maternal care, hippocampal synaptogenesis and cognitive development in rats," *Nat. Neurosci.*, vol. 3, pp. 799–806, 2000.

[12] E. M. Quinlan, B. D. Philpot, R. Huganir, , and M. Bear, "Rapid, experience-dependent expression of synaptic nmda receptors in visual cortex in vivo," *Nat. Neurosci.*, vol. 2, pp. 352–357, 1999.

[13] M. P. Mattson, W. Duan, and Z. Guo, "Meal size and frequency affect neuronal plasticity and vulnerability to disease: cellular and molecular mechanisms," *J. Neurochem.*, vol. 84, pp. 417–431, 2003.

[14] A. Wu, R. Molteni, Z. Ying, and F. Gomez-Pinilla, "A saturated-fat diet aggravates the outcome of traumatic brain injury on hippocampal plasticity and cognitive function by reducing brain-derived neurotrophic factor," *Neuroscience*, vol. 119, pp. 365–375, 2003.

[15] R. Molteni, R. J. Barnard, Z. Ying, C. K. Roberts, and F. Gomez-Pinilla, "A high-fat refined sugar diet reduces hippocampal brain-derived neurotrophic factor, neuronal plasticity, and learning," *Neuroscience*, vol. 112, pp. 803–814, 2002.

[16] W. Duan, J. Lee, Z. Guo, and M. Mattson, "Dietary restriction stimulates bdnf production in the brain andthereby protects neurons against excitotoxic injury," *Mol. Neurosci.*, vol. 16, pp. 1–12, 2001.

[17] H. van Praag, G. Kempermann, and F. Gage, "Neural consequences of environmental enrichment," *Nat. Rev. Neurosci.*, vol. 1, pp. 191–198, 2000.

[18] M. Rosenzweig and E. Bennett, "Psychobiology of plasticity: effects of training and experience on brain and behavior," *Behav. Brain Res*, vol. 78, pp. 57–65, 1996.

[19] E. L. Bennett, M. C. Diamond, D. Krech, and M. Rosenzweig, "Chemical and anatomical plasticity of brain," *Science*, vol. 164, pp. 610–619, 1964.

[20] M. C. Diamond, F. Law, H. Rhodes, B. Lindner, M. R. Rosenzweig, D. Krech, and E. L. Bennett, "Increases in cortical depth and glia numbers in rats subjected to enriched environment," *J. Comp. Neurol.*, vol. 128, pp. 117–126, 1966.

[21] M. R. Rosenzweig, W. Love, and E. L. Bennett, "Effects of a few hours a day of enriched experience on brain chemistry and brain weights," *Physiol Behav*, vol. 3, pp. 819–825, 1968.

[22] M. C. Diamond, C. A. Ingham, R. E. Johnson, E. L. Bennett, and M. R. Rosenzweig, "Effects of environment on morphology of rat cerebral cortex and hippocampus," *J. Neurobiol*, vol. 7, pp. 75–85, 1976.

[23] W. T. Greenough, F. R. Volkmar, and J. M. Juraska, "Effects of rearing complexity on dendritic branching in frontolateral and temporal cortex of the rat," *Exp. Neurol.*, vol. 41, pp. 371–378, 1973.

[24] A. M. Sirevaag, J. E. Black, D. Shafron, and W. Greenough, "Direct evidence that complex experience increases capillary branching and surface area in visual cortex of young rats," *Brain Res.*, vol. 471, pp. 299–304, 1988.

[25] M. C. Diamond, D. Krech, and M. R. Rosenzweig, "The effects of an enriched environment on the histology of the rat cerebral cortex," *J. Comp. Neurol.*, vol. 123, pp. 111–120, 1964.

[26] B. M. Williams, Y. Luo, C. Ward, K. Redd, R. Gibson, S. A. Kuczaj, and J. G. McCoy, "Environmental enrichment: effects on spatial memory and hippocampal creb immunoreactivity," *Physiol. Behav.*, vol. 73, pp. 649–658, 2001.

[27] R. C. Tees, K. Buhrmann, and J. Hanley, "The effect of early experience on water maze spatial learning and memory in rats," *Dev. Psychobiol.*, vol. 23, pp. 427–439, 1990.

[28] W. A. van Gool, M. Mirmiran, and F. van Haaren, "Spatial memory and visual evoked potentials in young and old rats after housing in an enriched environment," *Behav. Neural. Biol.*, vol. 44, pp. 454–469, 1985.

[29] R. A. Cummins, P. J. Livesey, and J. G. Evans, "A developmental theory of environmental enrichment," *Science*, vol. 197, pp. 692–694, 1977.

[30] H. van Praag, G. Kempermann, and F. H. Gage, "Neural consequences of environmental enrichment," *Neurosci. Nat. Rev.*, vol. 1, no. 3, pp. 191–198, Dec 2000.

[31] R. Chandler, S. Gibson, V. Karkare, S. Farshchi, D. Markovic, and J. Judy, "A system-level view of optimizing high-channel-count wireless biosignal telemetry," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE.* IEEE, 2009, pp. 5525–5530.

[32] S. Gibson, R. Chandler, V. Karkare, D. Markovic, and J. Judy, "An efficiency comparison of analog and digital spike detection," in *Neural Engineering, 2009. NER'09. 4th International IEEE/EMBS Conference on.* IEEE, 2009, pp. 423–428.

[33] P. Mohseni, K. Najafi, S. J. Eliades, and X. Wang, "Wireless multichannel biopotential recording using an integrated FM telemetry circuit," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 13, no. 3, pp. 263–271, 2005.

[34] P. Irazoqui-Pastor, I. Mody, and J. Judy, "Transcutaneous RF-powered neural recording device," *Proc. of the 24th Annual EMBS/BMES Conference, 2002*, vol. 3, 2002.

[35] 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. Solid-State Circuits*, vol. 42, no. 1, pp. 123–133, Jan. 2007.

[36] 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. Biomed. Eng.*, vol. 54, no. 1, pp. 130–137, 2007.

[37] C. Chestek, V. Gilja, P. Nuyujukian, S. Ryu, K. Shenoy, R. Kier, F. Solzbacher, and R. Harrison, "HermesC: RF wireless low-power neural recording system for freely behaving primates," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 1752–1755.

[38] H. Song, D. Allee, and K. Speed, "Single chip system for bio-data acquisition, digitization and telemetry," *Proceedings of 1997 IEEE International Symposium on Circuits and Systems*, vol. 3, 1997.

[39] A. Jackson, C. Moritz, J. Mavoori, T. Lucas, and E. Fetz, "The Neurochip BCI: towards a neural prosthesis for upper limb function." *IEEE Trans Neural Syst Rehabil Eng*, vol. 14, no. 2, pp. 187–90, 2006.

[40] I. Obeid, M. Nicolelis, and P. Wolf, "A multichannel telemetry system for single unit neural recordings," *J Neurosci Methods*, vol. 133, no. 1-2, pp. 33–8, 2004.

[41] M. Chae, W. Liu, Z. Yang, T. Chen, J. Kim, M. Sivaprakasam, and M. Yuce, "A 128-channel 6mW Wireless Neural Recording IC With On-the-fly Spike Sorting and UWB Transmitter," in *IEEE Int. Solid-State Circuits Conf. Dig.*, San Francisco, CA, USA, Feb 3–7, 2008, pp. 146–603.

[42] 31-channel and 63-channel Wireless Neural Headstage System. Triangle Biosystems, Inc. [Online]. Available: http://www.trianglebiosystems.com/Products/NeuralRecording.aspx

[43] A. M. Sodagar, K. D. Wise, and K. Najafi, "A Fully Integrated Mixed-signal Neural Processor for Implantable Multichannel Cortical Recording," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 6, pp. 1075–1088, 2007.

[44] R. R. Harrison, "A Low-Power Integrated Circuit for Adaptive Detection of Action Potentials in Noisy Signals," in *Proc. IEEE EMBS Conf.*, 2003, pp. 3325–3328.

[45] S. Gibson, J. W. Judy, and D. Marković, "Comparison of Spike-Sorting Algorithms for Future Hardware Implementation," in *Proc. 30th Ann. Int. Conf. IEEE EMBS*, Vancouver, British Columbia, Canada, Aug. 2008, pp. 5015–5020.

[46] R. Walker, H. Gao, P. Nuyujukian, K. Makinwa, K. Shenoy, T. Meng, and B. Murmann, "A 96-channel full data rate direct neural interface in 0.13μm cmos," in *VLSI Circuits (VLSIC), 2011 Symposium on*, june 2011, pp. 144 –145.

[47] R. R. Harrison, P. T. Watkins, R. J. Kier, D. J. Black, R. O. Lovejoy, R. A. Normann, and F. Solzbacher, "Design and Testing of an Integrated Circuit for Multi-electrode Neural Recording," in *VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on*, Bangalore, jan 2007, pp. 907–912.

[48] U. Rutishauser, E. Schuman, and A. Mamelak, "Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo," *Journal of neuroscience methods*, vol. 154, no. 1-2, pp. 204–224, 2006.

[49] R. R. Harrison and C. Charles, "A Low-power Low-noise CMOS Amplifier for Neural Recording Applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, 2003.

[50] C. Ferris, *Introduction to Bioinstrumentation With Biological, Environmental and Medical Applications*. Humana Pr Inc, 1979.

[51] G. Kovacs, "Introduction to the theory, design, and modeling of thin-film microelectrodes for neural interfaces," *Enabling Technologies for Cultured Neural Networks*, pp. 121–65, 1994.

[52] R. Plonsey and R. Barr, *Bioelectricity: a quantitative approach*. Springer, 2000.

[53] S. Gibson, J. Judy, and D. Markovic, "Technology-aware algorithm design for neural spike detection, feature extraction, and dimensionality reduction," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 18, no. 5, pp. 469 –478, 2010.

[54] T. Delbruck and C. A. Mead, "Adaptive Photoreceptor With Wide Dynamic Range," in *Circuits and Systems, 1994. ISCAS '94., 1994 IEEE International Symposium on*, vol. 4, London, 1994, pp. 339–342.

[55] Y. Perelman and R. Ginosar, "An Integrated System for Multichannel Neuronal Recording With Spike / LFP Separation and Digital Output," in *Neural Engineering, 2005. Conference Proceedings. 2nd International IEEE EMBS Conference on*, mar 16–19, 2005, pp. 377–380.

[56] R. H. I. 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. Biomed. Eng.*, vol. 52, no. 7, pp. 1303–1311, 2005.

[57] J. N. Y. Aziz, R. Genov, B. L. Bardakjian, M. Derchansky, and P. L. Carlen, "Brain-silicon Interface for High-resolution in Vitro Neural Recording," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, no. 1, pp. 56–62, 2007.

[58] B. Gosselin, M. Sawan, and C. A. Chapman, "A Low-power Integrated Bioamplifier With Active Low-frequency Suppression," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, no. 3, pp. 184–192, 2007.

[59] W. Wattanapanitch, M. Fee, and R. Sarpeshkar, "An Energy-efficient Micropower Neural Recording Amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, no. 2, pp. 136–147, 2007.

[60] T. Denison, K. Consoer, W. Santa, A. Avestruz, J. Cooley, and A. Kelly, "A 2 μW, 100nV/rtHz, Chopper-Stabilized Instrumentation Amplifier for Chronic Measurement of Neural Field Potentials," *IEEE J. Solid State Circuits*, vol. 42, no. 12, pp. 2934–2945, 2007.

[61] T. Borghi, A. Bonfanti, G. Zambra, R. Gusmeroli, A. S. Spinelli, and G. Baranauskas, "A Compact Multichannel System for Acquisition and Processing of Neural Signals," in *Int. Conf. IEEE Eng. in Medicine and Biology Soc.*, Lyon, aug 22–26, 2007, pp. 441–444.

[62] A. Agnes, E. Bonizzoni, P. Malcovati, and F. Maloberti, "A 9.4-ENOB 1V 3.8μW 100kS/s SAR ADC With Time-domain Comparator," in *IEEE Int. Solid-State Circuits Conf. Dig.*, San Francisco, CA, USA, 2008, pp. 246–610.

[63] V. Giannini, P. Nuzzo, V. Chironi, A. Baschirotto, G. V. der Plas, and J. Craninckx, "An 820μW 9b 40MS/s Noise-tolerant Dynamic-sar ADC in 90nm Digital CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig.*, San Francisco, CA, USA, Feb. 3–7, 2008, pp. 238–610.

[64] M. van Elzakker, E. van Tuijl, P. Geraedts, D. Schinkel, E. Klumperink, and B. Nauta, "A 1.9μW 4.4fJ/Conversion-step 10b 1MS/s Charge-Redistribution ADC," in *IEEE Int. Solid-State Circuits Conf. Dig.*, San Francisco, CA, 2008, pp. 244–610.

[65] P. T. Watkins, G. Santhanam, K. V. Shenoy, and R. R. Harrison, "Validation of Adaptive Threshold Spike Detector for Neural Recording," in *Proc. 26th Ann. Int. Conf. IEEE EMBS*, San Francisco, CA, USA, Sep. 2004, pp. 4079–4082.

[66] C. L. Rogers and J. G. Harris, "A Low-Power Analog Spike Detector for Extracellular Neural Recordings," in *Proc. 11th IEEE Int. Conf. Electronics, Circuits, and Systems*, Tel-Aviv, Israel, Dec. 2004, pp. 290–293.

[67] C. Rogers, J. Harris, J. Principe, and J. Sanchez, "An analog VLSI implementation of a multi-scale spike detection algorithm for extracellular neural recordings," in *Neural Engineering, 2005. Conference Proceedings. 2nd International IEEE EMBS Conference on*, March 2005, pp. 213–216.

[68] M. Chae, W. Liu, Z. Yang, T. Chen, J. Kim, M. Sivaprakasam, and M. Yuce, "A 128-Channel 6mW Wireless Neural Recording IC with On-the-Fly Spike Sorting and UWB Transmitter," in *IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, Feb. 2008, pp. 146–603.

[69] R. H. Olsson and K. D. Wise, "A Three-Dimensional Neural Recording Microsystem With Implantable Data Compression Circuitry," *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2796–2804, Dec. 2005.

[70] 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. Neural Eng.*, vol. 4, no. 3, pp. 309–321, Sep. 2007.

[71] I. Obeid and P. D. Wolf, "Evaluation of Spike-Detection Algorithms for a Brain-Machine Interface Application," vol. 51, no. 6, pp. 905–911, Jun. 2004.

[72] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '90)*, vol. 1, Albuquerque, NM, Apr. 1990, pp. 381–384.

[73] S. Mukhopadhyay and G. Ray, "A New Interpretation of Nonlinear Energy Operator and Its Efficacy in Spike Detection," vol. 45, no. 2, pp. 180–187, Feb. 1998.

[74] K. H. Kim and S. J. Kim, "Neural Spike Sorting Under Nearly 0-dB Signal-to-Noise Ratio Using Nonlinear Energy Operator and Artificial Neural-Network Classifier," vol. 47, no. 10, pp. 1406–1411, Oct. 2000.

[75] G. Anelli, "Design and Characterization of Radiation Tolerant Integrated Circuits in Deep Submicron CMOS Technologies for the LHC Experiments," Ph.D. dissertation, CERN, Grenoble, 2000.

[76] S. Farshchi, "An embedded system architecture for wireless neural recording," Ph.D. dissertation, University of California, Los Angeles, 2006.

[77] S. Farshchi, A. Pesterev, W. Ho, and J. W. Judy, "Acquiring high-rate neural spike data with hardware-constrained embedded sensors," in *Proc. of the 28th annual IEEE Engineering in Medicine and Biology Conference*, New York, NY, USA, Sept 2006.

[78] V. Karkare, S. Gibson, and D. Markovic, "A 130-μW, 64-channel neural spike-sorting DSP chip," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 5, pp. 1214 –1222, may 2011.

[79] M. Nicolelis, A. Ghazanfar, B. Faggin, S. Votaw, and L. Oliveira, "Reconstructing the Engram: Simultaneous, Multisite, Many Single Neuron Recordings," *NEURON*, vol. 18, pp. 529–537, 1997.

[80] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[81] Y. P. Zhang and Y. Hwang, "Time Delay Characteristics of 2.4 GHz Band Radio Propagation Channels in Room Environments," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, The Hague, Netherlands, Sep. 1994, pp. 28–32.

[82] D. Devasirvatham, "Multipath Time Delay Spread in the Digital Portable Radio Environment," vol. 25, no. 6, pp. 13–21, 1987.

[83] 31 and 63 Channel Wireless Neural Headstage System. Triangle Biosystems, Inc. [Online]. Available: http://www.trianglebiosystems.com/Products/NeuralRecording.aspx

[84] S. Chang and E. Yoon, "A 1μW 85nV$sqrt$Hz pseudo open-loop preamplifier with programmable band-pass filter for neural interface system," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 1631–1634.

[85] S. Rai, J. Holleman, J. Pandey, F. Zhang, and B. Otis, "A 500μW neural tag with 2μVrms AFE and frequency-multiplying MICS/ISM FSK transmitter," in *Solid-State Circuits Conference-Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*. IEEE, 2009, pp. 212–213.

[86] 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," *Journal of Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, 1995.

[87] D. Binkley, *Tradeoffs and Optimization in Analog CMOS Design*. Wiley Interscience, 2005.